5ª edição

com novos recursos

+ prática

EDITORA IBPEX DIALÓGICA

NELSON PEREIRA CASTANHEIRA

estatística

APLICADA A TODOS OS NÍVEIS

» ESTATÍSTICA APLICADA A TODOS OS NÍVEIS

O selo DIALÓGICA da Editora Ibpex faz referência às publicações que privilegiam uma linguagem na qual o autor dialoga com o leitor por meio de recursos textuais e visuais, o que torna o conteúdo muito mais dinâmico. São livros que criam um ambiente de interação com o leitor – seu universo cultural, social e de elaboração de conhecimentos –, possibilitando um real processo de interlocução para que a comunicação se efetive.

Nelson Pereira Castanheira

» ESTATÍSTICA APLICADA A TODOS OS NÍVEIS

EDITORA IBPEX DIALÓGICA

5ª edição revista e atualizada

EDITORA IBPEX
Av. Vicente Machado, 317 – 14º andar
Centro – CEP 80420-010 – Curitiba – PR – Brasil
Fone: (41) 2103-7306
www.editoraibpex.com.br
editora@editoraibpex.com.br

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Castanheira, Nelson Pereira.
Estatística aplicada a todos os níveis [livro eletrônico] / Nelson Pereira Castanheira. – 5. ed. rev. e atual.– Curitiba: Ibpex, 2012. – (Série Matemática aplicada).
2 Mb ; PDF

Bibliografia.
ISBN 978-85-7838-943-7

1. Estatística 2. Estatística matemática I. Título. II. Série.

12-08532 CDD-519.5

Índice para catálogo sistemático:
1. Estatística aplicada 519.5

Foi feito o depósito legal.

1ª edição, 2012.

Sumário

Dedico este livro à minha esposa Denise e aos meus filhos Kendric, Larissa, Marcella e Marcel, aos quais agradeço pela compreensão e colaboração durante a execução desta obra.

Apresentação

Este livro foi elaborado em capítulos e estruturado para permitir sua aplicação tanto em cursos presenciais quanto em cursos de educação a distância. É uma obra que possui uma linguagem de fácil compreensão, abordando os tópicos de forma progressiva e lógica, a fim de possibilitar um estudo agradável.

Nos capítulos 1 e 2, o autor faz uma apresentação da estatística e ensina como apresentar os dados obtidos em uma pesquisa.

No capítulo 3, é estudada exaustivamente a distribuição de frequências, ferramenta importantíssima para a tomada de decisões.

No capítulo 4, são apresentadas as medidas de tendência central, com ênfase na média, na mediana e na moda.

No capítulo 5, são estudadas as medidas de dispersão, em que o ponto alto é a determinação do desvio padrão de um conjunto de dados.

No capítulo 6, são realizados cálculos com as principais medidas de assimetria e de curtose, identificando onde elas se aplicam.

Nos capítulos de 7 a 10, é detalhadamente estudada a teoria das probabilidades, com especial atenção ao cálculo da distribuição de probabilidades.

No capítulo 11, são apresentados os passos da distribuição qui-quadrado, no 12 a obra nos mostra o estudo da inferência estatística e, no capítulo 13, como se efetua o teste de hipóteses. Ainda, acrescentamos o capítulo 14 com a análise de variância (Anova).

Boa leitura.

Conteúdos do capítulo

- Definição de estatística
- Finalidade dos métodos estatísticos
- Tipos de estatísticas: descritiva e indutiva
- Fases do método estatístico
- Fontes oficiais de consulta de dados

Após o estudo deste capítulo, você será capaz de:

1. desenvolver a lógica do cálculo porcentual em operações cotidianas;
2. aplicar o cálculo de transformação de uma razão qualquer em razão porcentual;
3. compreender o conceito de capitalização.

Logo na abertura do capítulo, você fica conhecendo os conteúdos que serão nele abordados. Você também é informado a respeito das competências que irá desenvolver e dos conhecimentos que irá adquirir com o estudo do capítulo.

Síntese

Adentrando no estudo da estatística descritiva, exercitamos e conceituamos neste capítulo aspectos relativos à distribuição de frequências, em que trabalhamos com população e amostra. Nessa construção de conhecimentos, especificamos as características das variáveis envolvidas no processo, as quais podem ser quantitativas ou qualitativas. Estas últimas, por sua vez, ainda podem ser classificadas em ordinal e nominal, enquanto as primeiras (as quantitativas) são inseridas em dois distintos grupos: as discretas e as contínuas. Como já dissemos, variáveis quantitativas são medições e contagens e as qualitativas são as usadas na descrição de aspectos de pertinência. Mas, para elaborar uma distribuição de frequência, também é necessário conhecer e saber determinar a quantidade de intervalos ou classes. E, como o objetivo é que você não apenas construa os conceitos, mas principalmente saiba aplicá-los, traçamos todo esse panorama através de exercícios práticos.

Você dispõe, ao final do capítulo, de uma síntese que traz os principais conceitos nele abordados.

Exercícios Resolvidos

1. Suponhamos que uma pesquisa resultou em 800 observações. Nesse caso, quantas classes ou intervalos temos?

Aplicando a fórmula do método de Sturges, temos que:

$i = 1 + 3{,}3 \cdot \log 800$

$i = 1 + 3{,}3 \cdot 2{,}9031$

$i = 10{,}58023$

Concluímos, portanto, que para 800 observações devemos ter 11 classes ou intervalos de valores.

Suponhamos, agora, uma pesquisa em que temos somente 40 observações: quantas são as classes (ou intervalos)?

$i = 1 + 3{,}3 \cdot \log 40$

$i = 1 + 3{,}3 \cdot 1{,}60206$

$i = 6{,}2868$

Resposta: No caso, temos sete classes (ou intervalos).

Durante o livro há uma série de exercícios resolvidos para facilitar a compreensão dos capítulos.

Questões para revisão

1. Ao realizarmos um teste de Estatística em uma turma constituída por 40 alunos, obtivemos os seguintes resultados (dados brutos):

   7 - 6 - 8 - 7 - 6 - 4 - 5 - 7 - 7 - 8 - 5 - 10 - 6 - 7 - 8 - 5 - 10 - 4 - 6 - 7 - 7 - 9 - 5 - 6 - 8 - 6 - 7 - 10 - 4 - 6 - 9 - 5 - 8 - 9 - 10 - 7 - 7 - 5 - 9 - 10.

   Qual o resultado que aconteceu com a maior frequência?

   ( ) 10 ( ) 6 ( ) 7

   ( ) 9 ( ) 5

   ( ) 8 ( ) 4

Com estas atividades, você tem a possibilidade de rever os principais conceitos analisados. Ao final do livro, o autor disponibiliza as respostas às questões, a fim de que você possa verificar como está sua aprendizagem.

capítulo 1

Introdução à estatística

Conteúdos do capítulo

- Definição de estatística
- Finalidade dos métodos estatísticos
- Tipos de estatísticas: descritiva e indutiva
- Fases do método estatístico
- Fontes oficiais de consulta de dados

Após o estudo deste capítulo, você será capaz de:

1. desenvolver a lógica do cálculo porcentual em operações cotidianas;
2. aplicar o cálculo de transformação de uma razão qualquer em razão porcentual;
3. compreender o conceito de capitalização.

O que é estatística?

É extremamente difícil definir estatística, e, tendo em vista que o seu domínio é muito amplo, o número de definições que encontramos é extremamente grande.

O dicionarista Aurélio Buarque de Holanda Ferreira[1] definiu-a como uma parte da matemática em que se investigam os processos de obtenção, organização e análise de dados sobre uma população ou sobre uma coleção de seres quaisquer, e os métodos de tirar conclusões e fazer predições com base nesses dados. Trata-se, portanto, de "uma metodologia desenvolvida para a coleta, a classificação, a apresentação, a análise e a interpretação de dados quantitativos e a utilização desses dados para a tomada de decisões"[2]

Devemos observar que esses dados se referem a fenômenos de massa, ou coletivos, e às relações que existem entre eles. É importante ressaltar que os dados – após análise – devem ser interpretados, uma vez que "as estatísticas 'mentem' apenas quando estão erradas ou, no mínimo, estão sendo mal interpretadas"[3]

> Parece difícil, não é? Mas não se preocupe. Você verá, ao longo deste estudo, que a estatística é simples.

Convém lembrar que, hoje, a parte "maçante" da estatística é realizada pelos computadores, e ao estudante ou ao profissional de estatística cabe interpretar e entender o significado do que se passa. Para a estatística,

somente interessam os fatos que englobam um grande número de elementos, pois ela busca encontrar leis de comportamento para todo o conjunto e não se preocupa com cada um dos elementos em particular.

As técnicas estatísticas, associadas a programas adequados de informática, constituem valiosos instrumentos para a administração.

A natureza dos métodos estatísticos

Métodos estatísticos são métodos para o tratamento de dados numéricos e referem-se a dados coletados, cujo destino é permitir que os estatísticos cheguem a conclusões sobre o que está sendo estudado (pessoas ou coisas).

População

Vimos anteriormente que a estatística tem por objetivo o estudo dos fenômenos de massa, ou coletivos, e das relações entre eles. Precisamos, portanto, ter bem claro que fenômeno coletivo é aquele que se refere a um grande número de elementos, sejam pessoas ou sejam coisas, aos quais denominamos de população ou universo. A estatística procura encontrar leis de comportamento para toda a população, ou universo; não se preocupa, portanto, com cada elemento em particular.

De acordo com o seu tamanho, a população, ou universo, pode ser classificada como finita ou infinita.

Exercícios Resolvidos

1. Se a POPULAÇÃO FINITA é aquela cujo número total (número finito) de elementos é conhecido, e estamos analisando o aproveitamento nas aulas de Estatística de uma turma de 50 alunos, sabemos, portanto, exatamente quantos alunos estão sendo observados. QUE TIPO DE POPULAÇÃO ESTAMOS OBSERVANDO?

   RESPOSTA: Uma população finita, pois a população de alunos é finita.

2. Também sabemos que a população que possuir um número infinito de elementos É INFINITA. Concorda? Portanto, se desejarmos saber quantas pétalas têm, em média, as rosas que florescem no Brasil, estaremos diante de uma situação em que não sabemos exatamente quantas são as rosas que florescem em nosso país. QUE TIPO DE POPULAÇÃO É A DESSAS ROSAS?

Resposta: A população de rosas é considerada infinita ou "praticamente" infinita. Lembre-se de que denominamos de população infinita a qualquer população cuja quantidade de elementos seja muito grande ou, então, seja difícil de ser quantificada.

Resumindo: População é o conjunto de elementos que desejamos observar para obtermos determinados dados.

Amostra

Quando a população é muito grande, certamente é difícil, ou mesmo impossível, a observação de determinada característica em todos os seus elementos. Daí a necessidade de selecionarmos uma parte finita dessa população, para que possamos realizar a observação e obter os dados que desejamos. A essa parte da população denominamos *amostra*.

Retomando o exemplo anterior: nele desejávamos saber quantas pétalas têm as rosas que florescem no Brasil. Como é impossível contar as pétalas de todas as rosas, selecionamos uma quantidade finita de rosas, 80 rosas, e contamos, uma a uma, as suas pétalas. Nesse exemplo estamos trabalhando com uma amostra (uma quantidade finita) de 80 rosas.

Resumindo: Amostra é o subconjunto de elementos retirados da população que estamos observando para obtermos determinados dados.

Estatística descritiva e estatística indutiva

A estatística descritiva, ou dedutiva, tem por objeto descrever e analisar determinada população, sem, com isso, pretender tirar conclusões de caráter mais genérico. É a parte da estatística referente à coleta e à tabulação dos dados.[4]

É comum o estatístico defrontar-se com a situação de dispor de tantos dados que se torna difícil absorver completamente a informação que está procurando investigar. É extremamente difícil captar intuitivamente todas as informações que os dados contêm. É necessário, portanto, que as informações sejam reduzidas até o ponto em que seja possível interpretá-las mais claramente.

A estatística descritiva é um número que, sozinho, descreve uma característica de um conjunto de dados, ou seja, é um número-resumo que possibilita reduzir os dados a proporções mais facilmente interpretáveis.[5]

A ESTATÍSTICA INDUTIVA, OU INFERÊNCIA ESTATÍSTICA, é a parte da estatística que, baseando-se em resultados obtidos da análise de uma amostra da população, procura inferir, induzir ou estimar as leis de comportamento da população da qual a amostra foi retirada. Refere-se a um processo de generalização a partir de resultados particulares; é, portanto, a parte da estatística concernente às conclusões sobre as fontes de dados.

Por exemplo, suponhamos que desejamos conhecer o grau de pureza de bauxita a ser exportada por um navio no porto de Santos. Como não podemos verificar esse grau de pureza em toda a população de bauxita, pegamos uma parte dessa população, como amostra, e procedemos, na mesma, os testes necessários e consideramos que o resultado obtido nesse teste é válido para toda a população de bauxita.

Esse processo de generalização, que é característico do método indutivo, está associado a uma margem de incerteza. A incerteza deve-se ao fato de a conclusão, que pretendemos obter para toda a população analisada, basear-se em uma amostra do total de observações. A medida da incerteza é tratada mediante técnicas e métodos que se fundamentam na "teoria das probabilidades".

É, então, importante você entender bem a definição de inferência estatística que resumiremos a seguir.

> INFERÊNCIA ESTATÍSTICA é admitirmos que os resultados obtidos na análise dos dados de uma amostra são válidos para toda a população da qual aquela amostra foi retirada. Consiste em obtermos e generalizarmos conclusões.

Fases do método estatístico (estatística descritiva)

Quando pretendemos realizar um estudo estatístico completo em determinada população ou em determinada amostra, o trabalho que realizaremos deve passar por várias fases, que são desenvolvidas até chegarmos aos resultados finais que procurávamos.

As principais fases são: definição do problema; delimitação do problema; planejamento para obtenção dos dados; coleta dos dados; apuração dos dados; apresentação dos dados; análise dos dados e interpretação dos dados.

Definição do problema

Consiste em definir com clareza o que pretendemos pesquisar, qual é o objeto de estudo e qual é exatamente o objetivo que desejamos alcançar.

Delimitação do problema

Não é suficiente saber com clareza o que pretendemos pesquisar. É também necessário saber onde será realizada a pesquisa: em que local, com que tipo de pessoas (ou coisas), em que dias (ou horários) e assim por diante.

Planejamento para obtenção dos dados

A fase seguinte é o planejamento, ou seja, respondemos às perguntas: como vamos fazer para resolver o problema? que dados serão necessários? como obter esses dados?

Às vezes, é suficiente a pura observação; no entanto, na maioria das ocasiões, é necessário elaborar um questionário ou um roteiro de entrevista. Nesse caso, são necessárias pessoas para distribuir questionários ou para realizar entrevistas. Aqui reside a maior preocupação do estatístico (ou pesquisador), que é conseguir a mão de obra com o perfil adequado a cada caso.

Ainda nessa fase deve estar bem claro o cronograma das atividades, bem como o tamanho da população ou da amostra a ser pesquisada e quanto pretendemos gastar para realizar a pesquisa.

Coleta dos dados

Essa fase consiste na obtenção de dados, propriamente ditos, seja por meio de simples observação, seja mediante a utilização de alguma ferramenta, como um questionário ou um roteiro de entrevista. É, provavelmente, a fase mais importante da pesquisa, pois se a forma utilizada não atender às expectativas, ocorre perda de tempo e de dinheiro.

Apuração dos dados

Antes de iniciarmos a apuração dos dados obtidos na pesquisa, devemos proceder à crítica dos mesmos, ou seja, descartar aqueles dados que foram fornecidos de forma errônea. Por exemplo, questionários respondidos pela metade não deverão ser levados em consideração. Nessa etapa resumimos os dados por meio de sua contagem, de separação por tipo de resposta e de agrupamento de dados semelhantes. É o que denominamos de *tabulação de dados*.

Apresentação dos dados

Os dados, uma vez apurados, podem ser apresentados em forma de tabelas ou em forma de gráficos.

Uma tabela consiste em dados dispostos em linhas e colunas distribuídas de modo ordenado, com a vantagem de exibir em um só local todos os resultados obtidos em determinada pesquisa, facilitando a possibilidade de análise e interpretação desses resultados.

Para facilitar ainda mais a visão do estatístico (ou pesquisador), podemos transformar em gráficos – cujos principais tipos você aprenderá mais adiante – os dados tabulados.

Análise dos dados

Nessa fase, o interesse principal do estatístico (ou pesquisador) é tirar conclusões que o auxiliem na solução do problema que o levou a executar a pesquisa. Tal análise está intimamente ligada ao cálculo de medidas que permite descrever, com detalhes, o fenômeno que está sendo analisado.

Interpretação dos dados

Para a interpretação dos dados analisados, devemos ter, em mãos, os dados tabulados, os gráficos (se tiverem sido feitos) e os cálculos das medidas estatísticas, que nos permitem até mesmo arriscar algumas generalizações. Lembramos que tais generalizações (a inferência estatística) são acompanhadas de um certo grau de incerteza, pois não podemos garantir cem por cento que os resultados obtidos numa amostra sejam totalmente verdadeiros para toda a população da qual aquela amostra pertence.

Órgãos e normas oficiais

Lembre-se: nem sempre você precisa fazer uma pesquisa para obter dados sobre determinado assunto. Às vezes, esses dados já existem e estão à disposição dos interessados em órgãos particulares ou governamentais; basta, para acessá-los, que tenhamos alguma dica de onde os encontrar.

Em geral, os dados nacionais (dados sobre o Brasil) podem ser obtidos junto ao Instituto Brasileiro de Geografia e Estatística (IBGE), que pode ser acessado pelo *site* <http://www. ibge.gov.br>.

Curitiba, por exemplo, possui o Instituto de Pesquisa e Planejamento Urbano de Curitiba (Ippuc), que pode ser acessado pelo *site* <http://www.ippuc.org.br> ou o Instituto Paranaense de Desenvolvimento Econômico e Social (Ipardes), que pode ser acessado pelo *site* <http://www. ipardes.gov.br>.

Outro bom exemplo é o Instituto Bonilha – Pesquisa de Opinião e Mercado S/C Ltda, http://www.bonilha.com.br ou o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), que pode ser acessado pelo *site* <http://www.inep.gov.br>.

Na verdade, temos uma infinidade de fontes de dados de pesquisas; os bancos, os ministérios, as bolsas de valores, os sindicatos e os tribunais são alguns exemplos.

Para dados internacionais, os órgãos também são muitos. Por exemplo, se desejamos informações sobre estatísticas sociais e trabalhistas, elas podem ser obtidas na Organização Internacional do Trabalho (OIT), com sede em Genebra, que pode ser acessado pelo *site* <http://www.ilo.org>.

Quanto às normas técnicas para apresentação tabular da estatística brasileira, há a Resolução nº 866 da Junta Executiva Central do Conselho Nacional de Estatística. A quem interessar, tais normas podem ser adquiridas junto a qualquer agência do IBGE (não é objetivo de nosso estudo o detalhamento e o entendimento de tais normas).

Síntese

Com o objetivo de situá-lo no universo da estatística, iniciamos este capítulo com a busca de uma possível definição para essa ciência, de modo que você possa compreender o valioso instrumento que ela representa para a administração. Também desenvolvemos a conceituação de "população" e "amostra", no âmbito estatístico. Fizemos isso por entendermos que é necessária uma visão clara sobre o que abrange o fenômeno estatístico – o coletivo. Sendo coletivo, nem sempre é possível de ser checado no todo, daí a importância de trabalharmos com amostras. Você ainda nos acompanhou na descrição das características da estatística descritiva e da estatística indutiva. Nesse contexto, verificamos que as fases do método estatístico compreendem as atividades de: definição do problema, delimi-

tação do problema, planejamento para a obtenção dos dados, seguido pela coleta, apuração, apresentação, análise e interpretação dos dados. No que se refere a obtenção de dados, importa salientar que muitas vezes basta acessá-los em órgãos governamentais ou particulares, sem que seja imprescindível a pesquisa de campo.

Questões para revisão

1. O que é população no âmbito da estatística? Elabore uma definição.

2. Defina o que é amostra para a estatística.

3. Assinale a assertiva que define o que é estatística descritiva.

   a) É o cálculo de medidas que permite descrever, com detalhes, o fenômeno que está sendo analisado.

   b) É a parte da estatística referente à coleta e à tabulação dos dados.

   c) É a parte da estatística referente às conclusões sobre as fontes de dados.

   d) É a generalização das conclusões sobre as fontes de dados.

   e) É a obtenção dos dados, seja por meio de simples observação ou mediante a utilização de alguma ferramenta.

4. Agora, assinale a assertiva que define a estatística indutiva.

   a) É o cálculo de medidas que permite descrever, com detalhes, o fenômeno que está sendo analisado.

   b) É a parte da estatística referente à coleta e à tabulação dos dados.

   c) É a parte da estatística referente às conclusões sobre as fontes de dados.

   d) É a generalização das conclusões sobre as fontes de dados.

   e) É a obtenção dos dados, seja por meio de simples observação ou mediante a utilização de alguma ferramenta.

5. Assinale com (F) se for falsa e com (V) se for verdadeira a afirmação. São duas fases do método estatístico:

    (   ) Criar um problema e coletar os dados.

    (   ) Criar um problema e analisar os dados.

    (   ) Planejar um problema e coletar os dados.

    (   ) Coletar os dados e analisar os dados.

    (   ) Apurar os dados e analisar um problema.

capítulo 2

Apresentação dos dados

Conteúdos do capítulo

- Tabulação e seus procedimentos
- A distribuição de frequências na apresentação de resultados
- A representação gráfica na apresentação de dados

Após o estudo deste capítulo, você será capaz de:

1. construir e identificar uma tabela estatística;
2. desenhar um gráfico e interpretá-lo;
3. dominar noções preliminares sobre distribuição de frequências e séries.

Tabulação

Dados brutos, rol e frequência

Dados brutos? Rol? Frequência? O que é isso?

É algo extremamente simples.

Imaginemos que um teste foi aplicado a uma turma de 50 alunos.

Lembre-se, esses 50 alunos constituem a população a ser pesquisada.

Imaginemos também que, à medida que faz a correção dos testes, o professor anote as notas obtidas pelos alunos. A transcrição desses resultados constitui o que denominamos, na estatística, de *dados brutos*.

> Logo dados brutos são a relação dos resultados obtidos em uma pesquisa e que foram transcritos aleatoriamente, ou seja, fora de qualquer ordem.

Os dados brutos são os dados originais, coletados em uma pesquisa, e que ainda não se encontram prontos para análise por não estarem numericamente organizados.[1]

Vamos supor que as notas do teste, a que nos referimos, tenham sido as seguintes:

7 - 6 - 8 - 9 - 6 - 5 - 7 - 4 - 6 - 8 - 9 - 8 - 7 - 6 - 10 - 8 - 4 - 5 - 6 - 10 - 5 - 8 - 4 - 3 - 8 - 7 - 9 - 6 -10 - 7 - 7 - 7 - 9 - 5 - 4 - 5 - 9 - 10 - 8 - 8 - 6 - 7 - 5 - 10 -8 - 6 - 7 - 7 - 10 - 6.

Em uma simples olhada para esses números, pouca ou nenhuma informação nos é transmitida. É preciso, então, primeiramente colocá-los em ordem, ou seja, transformá-los em um "rol".

Rol, portanto, é a relação dos resultados obtidos em uma pesquisa e que foram colocados em ordem numérica, crescente ou decrescente.

Coloquemos em ordem crescente os dados brutos anteriormente relacionados.

3 - 4 - 4 - 4 - 4 - 5 - 5 - 5 - 5 - 5 - 5 - 6 - 6 - 6 - 6 - 6 - 6 - 6 - 6 - 6 - 7- 7 - 7 - 7 - 7 - 7 - 7 - 7 - 7 - 7 - 8 - 8 - 8 - 8 - 8 - 8 - 8 - 8 - 8 - 9 - 9 - 9 - 9 - 9 - 10 - 10 - 10 - 10 - 10 - 10.

Melhorou! No entanto, ainda está confuso para tirarmos alguma conclusão. O que devemos fazer então?

Precisamos, agora, agrupar os valores que são iguais, para que possamos deduzir alguma coisa a respeito dos resultados obtidos nesse teste aplicado à turma de 50 alunos.

O primeiro passo consiste em verificarmos se temos notas que se repetem, isto é, se existem dois ou mais alunos com a mesma nota. Ao número de vezes que um mesmo valor se repete, denominamos de *frequência*.

Frequência ou frequência absoluta é o número de vezes que um mesmo resultado acontece durante uma pesquisa. Nós a denominaremos de ***f***.

No exemplo que estamos analisando, temos os resultados que estão na Tabela 1.

Tabela 1 – Notas obtidas pelos alunos

Notas	Frequência
3	1
4	4
5	6
6	9
7	10
8	9
9	5
10	6

Agora ficou mais fácil interpretar os resultados obtidos pela turma. Verificamos que a menor nota foi 3 e a maior foi 10. Verificamos ainda que a nota que ocorreu com a maior frequência (ou seja, que ocorreu o maior número de vezes) foi 7.

Frequência absoluta acumulada ou frequência acumulada (fa)

Frequência acumulada é o somatório das frequências dos valores inferiores ou iguais ao valor dado. Nós a designaremos por ***fa***. Vejamos, como exemplo, a Tabela 2.

Tabela 2 – Frequência acumulada da idade dos alunos

Idade	Frequência (f)	Frequência acumulada (fa)
4	4	= 4
5	5	+ = 9
6	7	+ = 16
7	4	+ = 20

Tabelas

Uma vez concluída a coleta e, também, a ordenação dos dados de uma pesquisa, devemos apresentá-los de tal forma que o leitor consiga identificar, rapidamente, uma série de informações. Para tal, a estatística costuma utilizar-se de duas ferramentas: tabelas e gráficos.

Vamos começar pelo estudo de tabelas.

A estrutura de uma tabela é constituída de três partes: cabeçalho, corpo e rodapé.[2]

O CABEÇALHO é a parte da tabela que contém o suficiente para esclarecer o leitor quanto ao que ela sintetiza. Por exemplo, NOTAS DA TURMA A EM ESTATÍSTICA – 1º BIMESTRE/2006.

O CORPO da tabela é constituído por linhas e colunas, nas quais são distribuídos os dados apurados na pesquisa.

O rodapé é o espaço no qual são colocadas as informações que permitem esclarecer a interpretação da tabela. Por exemplo, no rodapé colocamos a legenda e a fonte dos dados.

Exercício Resolvido

Vamos, então, representar em uma tabela os resultados do teste que foi aplicado à turma de 50 alunos, cujos resultados já foram ordenados anteriormente?

Tabela 3 – Notas da turma A em Estatística – 1º bimestre/2010

cabeçalho

Notas	Frequência
3	1
4	4
5	6
6	9
7	10
8	9
9	5
10	6

corpo

Fonte: Dados fictícios elaborados pelo autor.

rodapé

Distribuição de frequências

Quando estuda uma variável, o maior interesse do pesquisador é conhecer o comportamento dessa variável, analisando a ocorrência de suas possíveis realizações. Uma distribuição de frequências é a apresentação dos resultados de uma pesquisa por meio de uma tabela que mostra a frequência (o número de vezes) de ocorrência de cada resultado.

Assim, a Tabela 3, mostrada anteriormente, é uma distribuição de frequências.

Exercícios Resolvidos

Vamos praticar para que você tire suas dúvidas.

Vamos supor que desejamos conhecer a altura dos 50 alunos que realizaram o teste, anteriormente mencionado. Imaginemos que os resultados obtidos nas medições, em centímetros (dados brutos), foram:

160 - 168 - 174 - 169 - 180 - 165 - 173 - 174 - 170 - 168 - 172 - 164 - 164 - 170 - 167 - 178 - 175 - 165 - 176 - 165 - 161 - 170 - 172 - 179 - 160 - 177 - 174 - 167 - 162 - 171 - 177 - 167 - 163 - 177 - 165 - 178 - 170 - 170 - 174 - 165 - 166 - 170 - 170 - 168 - 178 - 166 - 175 - 176 - 168 - 167.

Agora, vamos transformar esses dados brutos em um ROL:

160 - 160 - 161 - 162 - 163 - 164 - 164 - 165 - 165 - 165 - 165 - 165 - 166 - 166 - 167 - 167 - 167 - 167 - 168 - 168 - 168 - 168 - 169 - 170 - 170 - 170 - 170 - 170 - 170 - 170 - 171 - 172 - 172 - 173 - 174 - 174 - 174 - 174 - 175 - 175 - 176 - 176 - 177 - 177 - 177 - 178 - 178 - 178 - 179 - 180.

Façamos agora uma tabela com a distribuição de frequências dessa pesquisa.

Tabela 4 – Alturas dos alunos da turma A – 1º bimestre/2010

Alturas (cm)	Frequência (f)
160	2
161	1
162	1
163	1
164	2
165	5
166	2
167	4
168	4
169	1
170	7
171	1
172	2
173	1
174	4
175	2
176	2
177	3
178	3
179	1
180	1

Fonte: Dados fictícios elaborados pelo autor.

Anote e guarde bem as definições de dados brutos, rol, frequência, frequência acumulada e de distribuição de frequências.

Classes ou intervalos

Quando o número de resultados obtidos em uma pesquisa é demasiadamente grande, é comum agruparmos esses resultados em faixas de valores, denominadas de *classes* ou *intervalos*. Por exemplo, se um pesquisador deseja saber a idade das pessoas pesquisadas, ele as distribui em faixas etárias.

No entanto, ao resumirmos os valores individuais em intervalos ou classes, estamos conscientes de que algum erro pode estar sendo inserido.

Lembram-se dos 50 alunos que fizeram o teste, cujos resultados tabulamos anteriormente? Suponhamos que estejamos interessados em saber a faixa etária em que se encontram esses alunos, conforme representado na tabela 5.

Tabela 5 – Faixa etária dos alunos

Faixa etária	Alunos (f)
0 a 5 anos	
6 a 10 anos	
11 a 15 anos	
16 a 20 anos	
21 a 25 anos	
26 a 30 anos	

Aos valores à esquerda de cada faixa etária, damos o nome de *limites inferiores* (Li), e aos valores à direita, damos o nome de *limites superiores* (Ls) das classes ou intervalos. A representação acima é utilizada pelos órgãos governamentais; contudo, em trabalhos acadêmicos, a forma de representar uma classe ou intervalo é a que mostramos na tabela 6.

Tabela 6 – Faixa etária dos alunos

Faixa etária	Alunos (f)
0 ⊢—— 5 anos	
5 ⊢—— 10 anos	
10 ⊢—— 15 anos	
15 ⊢—— 20 anos	
20 ⊢—— 25 anos	
25 ⊢—— 30 anos	

O símbolo ⊢—— representa que a classe ou intervalo é fechado à esquerda, ou seja, o valor escrito à esquerda (limite inferior) pertence ao intervalo e, como a classe ou intervalo é aberto à direita, o valor escrito à direita (limite superior) não pertence ao mesmo.

> Observe que, qualquer que seja a idade da pessoa pesquisada, ela se encaixa em apenas um dos intervalos.

Se subtrairmos o limite inferior do limite superior de determinada classe ou intervalo, temos o que se denomina de AMPLITUDE DO INTERVALO. Por exemplo, na tabela anterior, 25 menos 20 é igual a 5. Essa é a amplitude dos intervalos: 5.

Suponhamos que obtivemos o resultado abaixo, na pesquisa realizada.

Tabela 7 – Faixa etária dos alunos da turma A – 1º bimestre/2010

Faixa etária	Alunos (f)
0 ⊢—— 5 anos	0
5 ⊢—— 10 anos	6
10 ⊢—— 15 anos	14
15 ⊢—— 20 anos	20
20 ⊢—— 25 anos	8
25 ⊢——⊣ 30 anos	2

A distribuição dos dados em classes ou intervalos é comumente utilizada quando temos uma população muito grande para representar. Por exemplo, desejamos fazer uma tabela com os resultados dos 50.000 candidatos ao vestibular de determinada universidade federal; como as notas assumem, nesse caso, uma infinidade de valores, é conveniente agrupá-las em classes. A amplitude da classe quem define é o responsável pela tabela, entretanto, é oportuno observar que todas as classes devem ter a mesma amplitude, para facilitar os cálculos que iremos aprender adiante.

E QUANTAS CLASSES DEVE TER A TABELA?

Para que tenhamos confiabilidade nas informações, é apropriado que o número mínimo de classes ou intervalos seja igual a 5. Para que o visual da tabela não seja poluído, principalmente ao transformarmos a tabela em um gráfico, o número máximo de classes ou intervalos deve ser igual a 20. Lembre-se, esses números são apenas uma recomendação.

Outras grandezas são comumente agrupadas em classes ou intervalos, sendo que se destacam, entre elas, salários, pesos e alturas (entre outras).

Limites dos intervalos, ou classes

Esse assunto também já é de seu conhecimento. Mas vamos recordá-lo.

Os limites de um intervalo ou classe são os números extremos de cada intervalo ou classe. Por exemplo, na classe 20 ⊢—— 25, da tabela 7, o limite à esquerda (20) é denominado de *limite inferior* (Li) e o limite à direita (25) é denominado de *limite superior* (Ls).

No caso, o valor 20 pertence ao intervalo, ou classe, e o valor 25 não pertence.

> Lembre-se: um determinado valor só pode pertencer a um único intervalo (ou classe). No caso, o 25 pertence ao intervalo que vai de 25 a 30.

Amplitude dos intervalos, ou classes (A)

A amplitude do intervalo (ou classe) é obtida subtraindo-se o limite superior do limite inferior de qualquer classe da série (A = Ls – Li). Por exemplo, A = 25 – 20 = 5, então, na distribuição de frequências anterior, os intervalos (ou classes) têm amplitude A = 5.

Ponto médio do intervalo, ou classe (Pm)

A partir do momento que resolvemos agrupar os resultados obtidos numa pesquisa em intervalos (ou classes), assumimos que para todo intervalo o resultado é um valor único e igual ao ponto médio do respectivo intervalo. Por isso, afirmamos, anteriormente, que estamos conscientes de que um erro pode estar sendo inserido. Por exemplo, para o intervalo cujo limite inferior é 20 e cujo limite superior é 25, o ponto médio do intervalo é:

$$Pm = \frac{20 + 25}{2} = 22{,}5$$

Assumimos, desse modo, que as oito pessoas que estão nesse intervalo têm 22 anos e meio.

Agora, vamos agrupar em classes as alturas dos 50 alunos dos nossos problemas anteriores (ver Tabela 4); para isso faremos intervalos com amplitude igual a 3 centímetros (ver Tabela 8).

No capítulo 3, estudaremos, em detalhes, como determinar esse número de classes ou intervalos.

Tabela 8 – Alturas dos alunos da turma A – 1º bimestre/2006

Alturas	Alunos (f)
160 ⊢—— 163	4
163 ⊢—— 166	8
166 ⊢—— 169	10
169 ⊢—— 172	9
172 ⊢—— 175	7
175 ⊢—— 178	7
178 ⊢—— 181	5

Frequência relativa (fr)

A frequência relativa de um valor é dada pela fórmula:

$fr = \frac{f}{n}$, em que $n = \Sigma f$

Para exemplificar, vamos analisar a distribuição de frequências da tabela 9.

Tabela 9 – Frequência relativa das idades dos alunos

Idades	Frequência(f)	Frequência relativa (fr)
18 ⊢—— 21	9	9/100
21 ⊢—— 24	12	12/100
24 ⊢—— 27	12	12/100
27 ⊢—— 30	17	17/100
30 ⊢—— 33	16	16/100
33 ⊢—— 36	14	14/100
36 ⊢—— 39	11	11/100
39 ⊢——⊣ 42	9	9/100
Σ*	100	100/100 = 1

* O símbolo Σ é a letra grega maiúscula *sigma*, correspondente ao S do nosso alfabeto. Aqui, Σ significa somatório e Σf significa o somatório de f, isto é, $f_1 + f_2 + ... f_n$.

Séries estatísticas

Série estatística é a denominação que se dá a uma tabela na qual há um critério distinto que a especifica e a diferencia. Assim, podemos classificar as séries estatísticas em:

a) temporais (ou cronológicas, ou evolutivas, ou históricas);

b) geográficas (ou de localização, ou territorial, ou espacial);

c) específicas (ou categóricas, ou de qualidade);

d) conjugadas (ou mistas);

e) de distribuição de frequências.

As séries estatísticas diferenciam-se de acordo com a variação de um dos seguintes elementos: tempo (época), local (fator geográfico) e fato (fenômeno).

Séries estatísticas temporais

A série temporal tem como característica a variação do tempo (época), enquanto o local (fator geográfico) e o fato (fenômeno) permanecem fixos. São, portanto, séries em que os dados são produzidos (observados) ao longo do tempo. Ver Tabela 10.

Tabela 10 – Número de veículos exportados no mês de janeiro, pelo Brasil, nos últimos cinco anos, da marca X

Ano	Automóveis exportados
2006	998
2007	1.234
2008	1.405
2009	890
2010	1.665

Séries estatísticas geográficas

A série geográfica tem como característica a variação do local de ocorrência (fator geográfico), enquanto o tempo (a época) e o fato (o fenômeno) permanecem fixos. Ver Tabela 11.

Tabela 11 – Estimativas populacionais do Brasil – grandes regiões

Região	População
Norte	12.900.704
Nordeste	47.741.711
Sudeste	72.412.411
Sul	25.107.616
Centro-Oeste	11.636.728

Fonte: IBGE, Censo Demográfico 2000.

Séries estatísticas específicas

A SÉRIE ESPECÍFICA tem como característica a variação do fato (variação do fenômeno), enquanto que o local (fator geográfico) e o tempo (a época) são constantes. Isso significa que os dados são agrupados segundo a modalidade de ocorrência. Ver Tabela 12.

Tabela 12 – Número de candidatos para o vestibular em uma determinada universidade – 2010 (terceiro grau)

Áreas ofertadas	Número de candidatos
Ciências Sociais Aplicadas	5.404
Ciências Exatas	3.221
Ciências Humanas	6.558
Ciências Biológicas	2.489
Ciências Tecnológicas	1.559

Observe que, nesse caso, o fato se apresenta como um todo. Caso o fato se apresente disposto em classes, ou intervalos (dados agrupados), dá origem à chamada *distribuição de frequências*, já estudada por nós.

Séries estatísticas conjugadas

Uma SÉRIE É CONJUGADA, ou mista, quando existe a combinação entre as séries temporais, geográficas e específicas. Podem, portanto, variar o tempo (a época), o local (fator geográfico) e o fato (o fenômeno) simultaneamente. Ver tabela 13.

Tabela 13 – Número de veículos zero km da marca X vendidos no Brasil, nos últimos cinco anos, por região

Veículos vendidos					
Ano	Norte	Nordeste	Sudeste	Sul	Centro-Oeste
2006	1.845	4.111	1.3802	8.901	3.470
2007	1.903	4.100	1.3944	9.380	3.401
2008	1.899	4.189	1.3784	9.182	3.532
2009	1.847	4.333	1.4053	9.044	3.744
2010	2.005	4.704	1.4188	10.085	3.987

A distribuição de frequências, tal como a estudamos, é considerada uma quinta modalidade de série estatística. Trata-se de uma série estatística específica, em que os dados estão dispostos em classes, com suas respectivas frequências absolutas. Ver Tabela 14.

Tabela 14 – Quantidade de pessoas, por faixa etária, que têm plano de saúde em determinada localidade

Idade	N° de pessoas
0 ⊢ 10	208
10 ⊢ 20	431
20 ⊢ 30	644
30 ⊢ 40	955
40 ⊢ 50	938
50 ⊢ 60	745
60 ⊢ 70	730
70 ⊢⊣ 80	433

Gráficos

A representação gráfica é um complemento da apresentação dos dados em forma de tabelas, uma vez que permite uma rápida visualização do fato estudado. Todo gráfico, assim como as tabelas, tem na sua parte superior um título, e na parte inferior é dada a fonte que forneceu o gráfico ou os dados que permitiram a sua construção.

Há uma infinidade de gráficos. Vamos verificar os principais tipos a serem utilizados na estatística, ou seja, gráfico de colunas, de barras, de setores, bem como histograma e polígono de frequências.

Gráfico de colunas

O gráfico de colunas é utilizado para as séries temporais, geográficas e específicas. Como construí-lo?

Primeiro, trace um par de eixos ortogonais (sistema de eixos cartesianos):

a) o eixo horizontal (eixo x) chama-se *eixo das abscissas* e a sua escala cresce da esquerda para a direita, a partir da origem (interseção dos eixos horizontal e vertical);

b) o eixo vertical (eixo y) chama-se *eixo das ordenadas* e a sua escala cresce de baixo para cima, a partir da origem.

Na sequência, apresente a variável que está sendo estudada no eixo das abscissas e a frequência no eixo das ordenadas; após o que, dê um título ao gráfico.

Exercício Resolvido

Vamos transformar a tabela 10, apresentada nas séries estatísticas temporais, no gráfico 1?

Gráfico 1 – Número de veículos exportados no mês de janeiro, pelo Brasil, nos últimos cinco anos, da marca X

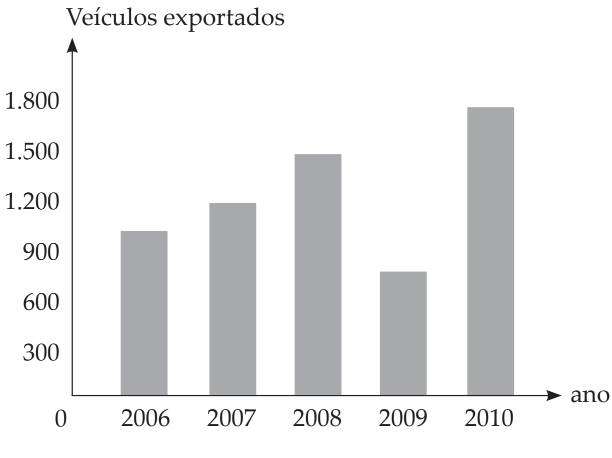


Gráfico de barras

É semelhante ao anterior, porém os retângulos são dispostos na horizontal. É também utilizado para as séries temporais, geográficas e específicas.

Exercício Resolvido

Agora, vamos transformar a Tabela 11, apresentada nas séries estatísticas geográficas, no Gráfico 2?

Gráfico 2 – Estimativas populacionais do Brasil – grandes regiões

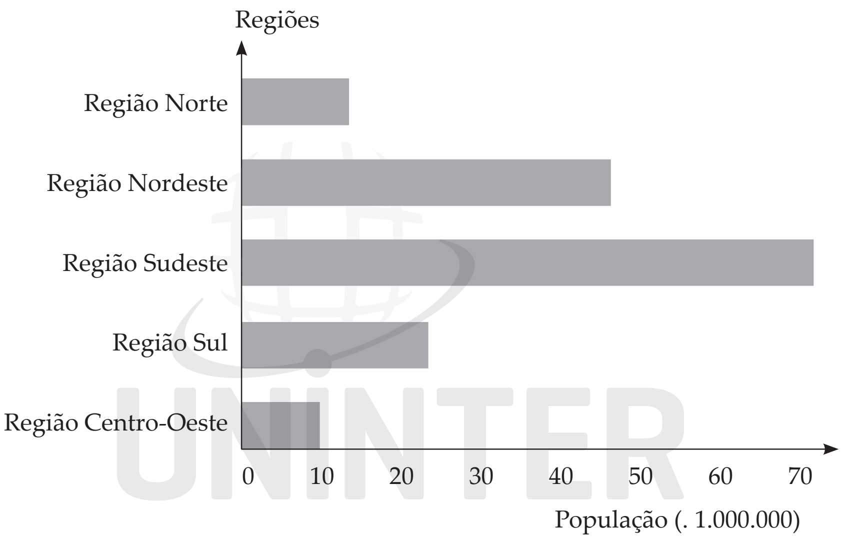


Gráfico de setores

O gráfico de setores é popularmente conhecido como *gráfico em forma de pizza*. Trata-se de uma representação em um círculo dividido em setores. É muito útil quando desejamos comparar cada valor da série estatística com o total.

Como construí-lo?

É só seguir os passos que detalharemos na sequência:

a) trace uma circunferência (lembre-se que a área do círculo corresponde ao todo, ou seja, 100%);

b) divida o círculo em setores, cujas áreas são proporcionais aos valores da série estatística (essa divisão pode ser obtida pela solução de uma regra de três simples), ou seja:

total ⟶ 360°

parte ⟶ x°

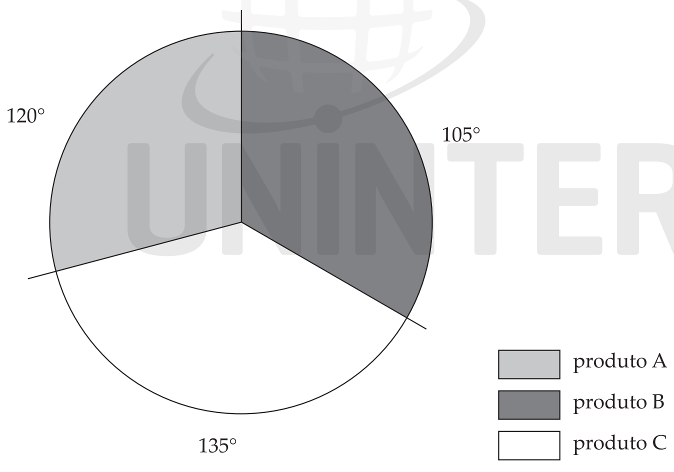
c) preencha cada setor de modo diferente ou com cores diferentes;

d) dê um título ao gráfico;

e) faça uma legenda para explicar os preenchimentos ou as cores utilizadas.

Observe um gráfico de setores no Gráfico 3.

Gráfico 3 – Exportações dos produtos A, B, C no mês de setembro de 2006



Histograma

O histograma é um gráfico formado por um conjunto de retângulos justapostos e é muito utilizado para representar a distribuição de frequências cujos dados foram agrupados em classes ou intervalos de mesma amplitude.

Como construí-lo?

É só observar as seguintes etapas:

a) trace um par de eixos ortogonais (sistema de eixos cartesianos);

b) no eixo horizontal (eixo x), anote os valores individuais da variável em estudo, ou seja, os valores dos intervalos das classes;

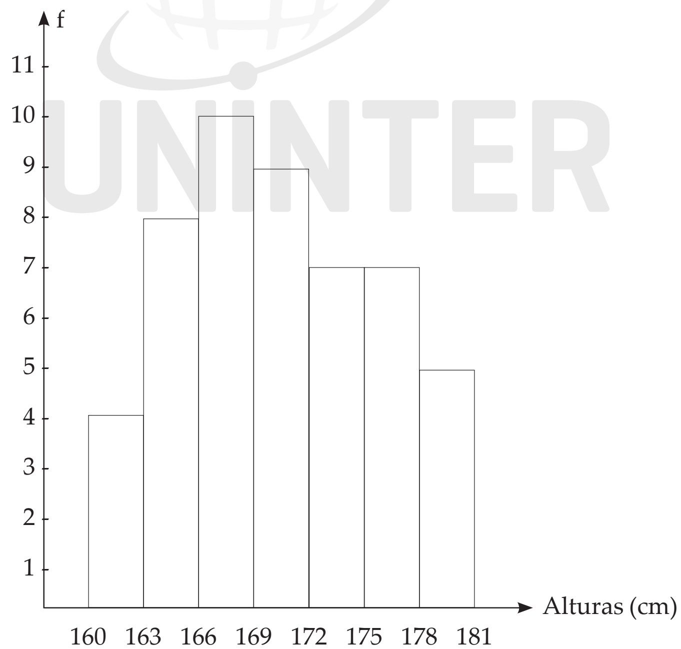
c) no eixo vertical (eixo y), anote os valores dos números de observações, ou seja, as frequências das classes;

d) dê um título ao gráfico.

Exercício Resolvido

E como transformarmos uma tabela em um histograma? Vamos, então, transformar a Tabela 8, apresentada nas séries estatísticas de distribuição de frequências no Gráfico 4, ou seja, em um histograma?

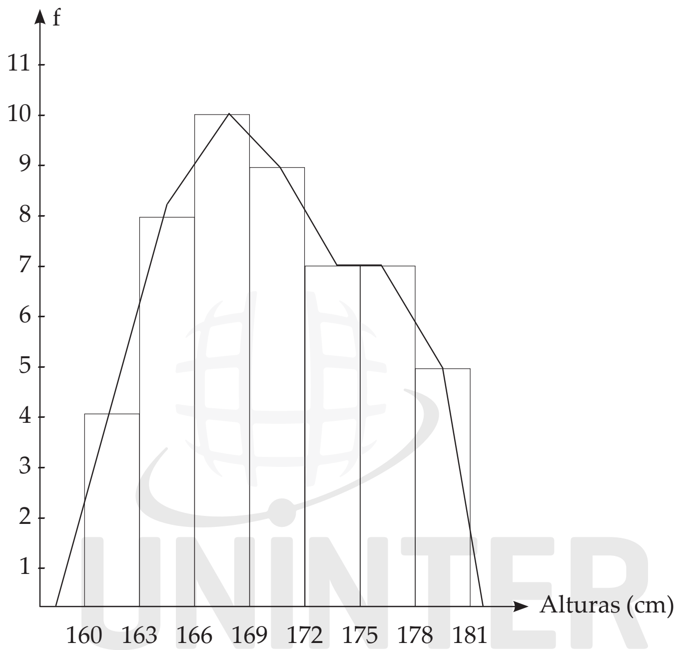
Gráfico 4 – Alturas dos alunos da turma A – 1º bimestre/2010



Polígono de frequências

O polígono de frequências é obtido unindo-se por segmentos de reta os pontos médios das bases superiores dos retângulos de um histograma. Como exemplo, vamos transformar o gráfico anterior (Gráfico 4) em um polígono de frequências (Gráfico 5).

Gráfico 5 – Alturas dos alunos da turma A – 1º bimestre/2010



Síntese

Se pretendermos ser capazes de construir e identificar uma tabela estatística, precisamos construir antes algumas concepções por meio do desenvolvimento de determinadas habilidades. No caso, precisamos conhecer fatores como dados brutos, rol e frequência para sermos capazes de realizar uma tabulação. No entanto, a realização de uma tabulação também requer conhecimentos de frequência acumulada (fa), habilidade na estruturação de tabelas, bem como na distribuição de frequências. Mas, para desenhar um gráfico e interpretá-lo, outras habilidades devem ser somadas àquelas, como identificar as séries estatísticas. Esse foi o processo que procuramos desenvolver neste capítulo.

Questões para revisão

1. Ao realizarmos um teste de Estatística em uma turma constituída por 40 alunos, obtivemos os seguintes resultados (dados brutos):

   7 - 6 - 8 - 7 - 6 - 4 - 5 - 7 - 7 - 8 - 5 - 10 - 6 - 7 - 8 - 5 - 10 - 4 - 6 - 7 - 7 - 9 - 5 - 6 - 8 - 6 - 7 - 10 - 4 - 6 - 9 - 5 - 8 - 9 - 10 - 7 - 7 - 5 - 9 - 10.

   Qual o resultado que aconteceu com a maior frequência?

   ( ) 10 ( ) 6 ( ) 7

   ( ) 9 ( ) 5

   ( ) 8 ( ) 4

2. Observe a tabela a seguir, e depois marque a alternativa correta na relação apresentada no começo da próxima página.

Ano	Exportações (em US$ 1.000.000,00)
2006	204
2007	234
2008	652
2009	888
2010	1.205

Fonte: Dados fictícios elaborados pelo autor.

A série estatística representada é:

( ) Cronológica.

( ) Geográfica.

( ) Conjugada.

( ) Específica.

( ) Espacial.

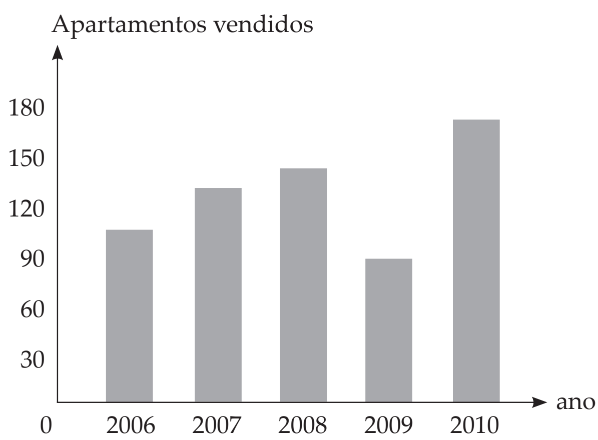
3. Assinale a resposta correta. Na distribuição de frequências a seguir, qual a amplitude das classes ou intervalos?

Ano	Alunos (f)
20 ⊢—— 25	8
25 ⊢—— 30	8
30 ⊢—— 35	8
35 ⊢—— 40	8
40 ⊢—— 45	8
45 ⊢—— 50	8

Fonte: Dados fictícios elaborados pelo autor.

( ) 30 ( ) 8 ( ) 50 ( ) 5 ( ) 6

4. Observe o gráfico representado a seguir.



Nós vimos que os gráficos podem ser de setores, de barras, de colunas, em forma de histograma ou de polígono. Como você classifica este que foi aqui representado?

5. Descreva quais são as partes que constituem uma tabela.

capítulo 3

Distribuição de frequências

Conteúdos do capítulo

- A finalidade e operacionalidade da distribuição de frequências.
- Classificação e contextualização das variáveis qualitativas e quantitativas.
- Determinação do número de classes ou intervalos.

Após o estudo deste capítulo, você será capaz de:

1. identificar todas as partes constituintes de uma tabela representativa de uma distribuição de frequências;
2. interpretar os dados de uma tabela representativa de uma distribuição de frequências.

A distribuição de frequências

No capítulo anterior estudamos séries estatísticas. Dos tipos estudados, a mais importante na estatística descritiva é a distribuição de frequências. Havíamos definido que:

> SÉRIE ESTATÍSTICA é a denominação que se dá a uma tabela na qual há um critério distinto que a especifica e a diferencia;
> A DISTRIBUIÇÃO DE FREQUÊNCIAS é uma série estatística específica, em que os dados estão dispostos em classes, com suas respectivas frequências absolutas.

Vamos rever alguns conceitos e aprender outros.

População

> População é o conjunto de elementos que desejamos observar para obter determinados dados.

A população pode conter um número conhecido de elementos e, nesse caso, a denominamos de *população finita*. Por exemplo, podemos estar interessados em estudar o consumo individual de água em 40 residências de um condomínio fechado, portanto sabemos exatamente quantas são as residências a analisar, ou seja, 40.

No entanto, nem sempre é assim, a população pode conter um número desconhecido ou um número muito grande de elementos, caso este em que denominamos de *infinita*; podemos, por exemplo, estar interessados em saber

a intenção de voto em determinado candidato à presidência da república – uma população infinita.

Amostra

> Amostra é o subconjunto de elementos retirados da população que estamos observando para obter determinados dados.

Variáveis

Podemos definir a variável como uma característica que observamos numa pesquisa e que pode assumir diferentes valores.

Por exemplo, se numa pesquisa estivermos interessados em anotar o sexo das pessoas que dela participam, podemos associar os valores masculino ou feminino à variável sexo.

Observamos que algumas variáveis descrevem qualidades (ou categorias, ou atributos) tais como sexo, escolaridade, *status* social, estado civil e tipo sanguíneo, entre outros. Tais variáveis são denominadas *qualitativas* e, normalmente, não podem ser expressas em valores numéricos.

> As variáveis qualitativas podem ainda ser classificadas em dois diferentes grupos: nominais e ordinais.

A VARIÁVEL QUALITATIVA NOMINAL permite somente a classificação dos dados, como é o caso da variável sexo e do ramo de atividade de uma empresa, entre outras.

A VARIÁVEL QUALITATIVA ORDINAL permite que se estabeleça uma ordem nos seus resultados como, por exemplo, o grau de instrução ou o *status* (classe) social de um grupo de pessoas.

Observamos, entretanto, que algumas variáveis podem ser expressas por meio de valores numéricos, como número de filhos ou de dependentes, idade, peso (massa), altura, quantidade de sacas de milho colhidas por hectare, entre outros. Essas variáveis são denominadas de *quantitativas*.

> As variáveis quantitativas também podem ser classificadas em dois diferentes grupos: discretas e contínuas.

A VARIÁVEL QUANTITATIVA DISCRETA permite relacionar todos os possíveis valores que ela pode assumir. Além disso, apresenta lacunas entre os valores que pode tomar para si, tais como número de peças defeituosas produzidas por determinada máquina ou o número de filhos dos empregados de determinada empresa.

A VARIÁVEL QUANTITATIVA CONTÍNUA, por sua vez, pode assumir infinitos valores dentro de um intervalo de números reais, de tal forma que não podemos previamente relacionar todos os possíveis resultados a encontrar na pesquisa. Como exemplo, podemos citar a altura (estatura) dos empregados de uma fábrica ou as diferentes temperaturas registradas ao longo de certo tempo em uma localidade.

Às vezes, podemos atribuir valores numéricos a determinada variável qualitativa e tratá-la como se fosse quantitativa. Quando isso é possível? Quando, a essa variável qualitativa, atribuirmos somente dois possíveis valores. Você verá, no capítulo 8, que esses dois valores serão denominados de *sucesso* e *insucesso* (ou fracasso).

> Resumidamente, VARIÁVEIS QUANTITATIVAS são medições e contagens. VARIÁVEIS QUALITATIVAS descrevem pertinência ao grupo.
> Variáveis com poucos valores repetidos são consideradas contínuas. Variáveis com muitos valores repetidos são consideradas discretas.[1]

Convém ressaltar que, caso a variável em estudo seja contínua, é conveniente agrupar os valores obtidos em classes ou intervalos.

Número de classes ou intervalos

COMO DETERMINAR A QUANTIDADE DE INTERVALOS EM UMA DISTRIBUIÇÃO DE FREQUÊNCIAS?

Para tal, não há fórmula exata. Entretanto, recomenda-se que o número mínimo de intervalos seja igual a 5 e o número máximo igual a 20, o que facilita a construção da tabela e do respectivo gráfico, com um mínimo de precisão e de informação.

Como exemplo, suponhamos que perguntamos a idade a um grupo de 100 pessoas e, ainda, que desejamos representar os resultados obtidos distribuídos em intervalos (ou classes). Vamos presumir os seguintes resultados, já dispostos em forma de rol:

18 - 18 - 18 - 19 - 19 - 19 - 19 - 20 - 20 - 21 - 21 - 21 - 21 - 22 - 22 - 22 - 23 - 23 - 23 - 23 - 23 - 24 - 24 - 24 - 24 - 25 - 25 - 25 - 26 - 26 - 26 - 26 - 26 - 27 - 27 - 27 - 28 - 28 - 28 - 28 - 28 - 29 - 29 - 29 - 29 - 29 - 29 - 29 - 29 - 29 - 30 - 30 - 30 - 30 - 31 - 31 - 31 - 31 - 31 - 31 - 32 - 32 - 32 - 32 - 32 - 32 - 33 - 33 - 33 - 33 - 33 - 33 - 34 - 34 - 34 - 34 - 35 - 35 - 35 - 35 - 36 - 36 - 36 - 37 - 37 - 37 - 37 - 38 - 38 - 38 - 38 - 39 - 39 - 40 - 40 - 40 - 41 - 41 - 42 - 42.

Nesse caso, inicialmente elaboramos uma tabela de distribuição de frequências, considerando-se os valores individuais.

Tabela 15 – Distribuição das idades de 100 pessoas

Idade	Frequência (f)
18	3
19	4
20	2
21	4
22	3
23	5
24	4
25	3
26	5
27	3
28	5
29	9
30	4
31	6
32	6
33	6
34	4
35	4
36	3
37	4
38	4
39	2
40	3
41	2
42	2

Na sequência, agrupamos esses resultados em intervalos (ou classes); para tal devemos definir quantos intervalos queremos fazer, tomando o cuidado para que todos tenham o mesmo tamanho (ou amplitude). Podemos, por exemplo, optar por oito intervalos de amplitude igual a 3.

Mas, como obtivemos esses valores? Observe que a idade maior é 42 anos e a menor é 18 anos. Logo, temos o intervalo de:

42 – 18 = 24 (essa é amplitude total, também denominada de *range*)

24 : 8 = 3

Teremos, então, a tabela 16.

Tabela 16 – Distribuição das idades em intervalos

Ano	Alunos (f)
18 ⊢ 21	9
21 ⊢ 24	12
24 ⊢ 27	12
27 ⊢ 30	17
30 ⊢ 33	16
33 ⊢ 36	14
36 ⊢ 39	11
39 ⊢⊣ 42	9

A representação ⊢⊣ indica que tanto o 39 quanto o 42 pertencem a essa classe ou intervalo.

Para a utilização de uma fórmula que nos permita determinar a quantidade de intervalos, devemos tomar o cuidado para que todos tenham a mesma amplitude (A). Você lembra que recomendamos que esse número esteja entre os valores 5 e 20. Já vimos, também, que A = Ls – Li.

Chamamos de i o número de classes ou intervalos. Então, a amplitude de cada classe será igual a:

$$A = \frac{A_t}{i}$$

Caso essa divisão resulte em um valor não inteiro, convém fazer um arredondamento para o número inteiro mais próximo, sempre para um número maior que o resultado encontrado na divisão.

Como sabemos, com o número de classes (i) e a amplitude de cada classe (A) podemos construir a tabela de distribuição de frequências.

Salientamos o fato de que o número de classes ou intervalos pode ser determinado por diversos métodos, entre os quais se destaca o método de *Sturges*, em que:

$i = 1 + 3{,}3 \cdot \log n$

Sendo n o número total de observações, ou seja, a quantidade de dados que temos no rol.

Exercícios Resolvidos

1. Suponhamos que uma pesquisa resultou em 800 observações. Nesse caso, quantas classes ou intervalos temos?

Aplicando a fórmula do método de Sturges, temos que:

$i = 1 + 3{,}3 \cdot \log 800$

$i = 1 + 3{,}3 \cdot 2{,}9031$

$i = 10{,}58023$

Concluímos, portanto, que para 800 observações devemos ter 11 classes ou intervalos de valores.

Suponhamos, agora, uma pesquisa em que temos somente 40 observações: quantas são as classes (ou intervalos)?

$i = 1 + 3{,}3 \cdot \log 40$

$i = 1 + 3{,}3 \cdot 1{,}60206$

$i = 6{,}2868$

Resposta: No caso, temos sete classes (ou intervalos).

Entretanto, na prática, costumamos usar o bom senso na hora de determinar essa quantidade e não utilizamos qualquer método ou regra que aplique uma fórmula.

2. As notas obtidas pelos alunos de uma classe, em um teste de Estatística, estão descritas a seguir em uma tabela de distribuição de frequências.

Tabela 17 – Notas obtidas em um teste de Estatística distribuídas em intervalos

Notas	Frequência (f)
0 ⊢ 1	4
1 ⊢ 2	5
2 ⊢ 3	7
3 ⊢ 4	8
4 ⊢ 5	9
5 ⊢ 6	14
6 ⊢ 7	8
7 ⊢ 8	6
8 ⊢ 9	5
9 ⊢⊣ 10	4
Σ	70

Em relação a essa tabela, pedimos para determinar:

a) a amplitude total;

Para o cálculo da amplitude total, subtraimos o maior valor do menor dos resultados obtidos.

No caso: 10 – 0 = 10.

b) a quantidade de classes ou intervalos;

Basta contar quantos são os intervalos; no caso, 10.

c) a amplitude das classes ou intervalos;

Subtraimos o limite superior de uma classe qualquer do seu limite inferior. Por exemplo, na quinta classe, 5 – 4 = 1; então, as classes têm amplitude igual a 1.

d) o limite superior da terceira classe;

As classes são contadas de cima para baixo; logo, a terceira classe compreende as notas de 2 até 3; sendo que o limite superior dessa classe é o 3.

e) o limite inferior da quinta classe;

O limite inferior é o 4, pois a quinta classe compreende as notas de 4 até 5.

f) O PONTO MÉDIO DA DÉCIMA CLASSE;

A décima classe compreende os valores de 9 até 10; então, o ponto médio é

$$\frac{10+9}{2} = 9{,}5$$

g) A FREQUÊNCIA DA SEGUNDA CLASSE;

Basta consultar a tabela; a segunda classe tem frequência igual a 5.

h) A FREQUÊNCIA ACUMULADA ATÉ A QUINTA CLASSE;

Devemos somar as frequências das primeira, segunda, terceira, quarta e quinta classes.

Assim, obtemos a frequência acumulada de:

$4 + 5 + 7 + 8 + 9 = 33$.

i) A FREQUÊNCIA RELATIVA DA NONA CLASSE OU INTERVALO.

$$fr = \frac{5}{70} = \frac{1}{14}$$

Síntese

Adentrando no estudo da estatística descritiva, exercitamos e conceituamos neste capítulo aspectos relativos à distribuição de frequências, em que trabalhamos com população e amostra. Nessa construção de conhecimentos, especificamos as características das variáveis envolvidas no processo, as quais podem ser quantitativas ou qualitativas. Estas últimas, por sua vez, ainda podem ser classificadas em ordinal e nominal, enquanto as primeiras (as quantitativas) são inseridas em dois distintos grupos: as discretas e as contínuas. Como já dissemos, variáveis quantitativas são medições e contagens e as qualitativas são as usadas na descrição de aspectos de pertinência. Mas, para elaborar uma distribuição de frequência, também é necessário conhecer e saber determinar a quantidade de intervalos ou classes. E, como o objetivo é que você não apenas construa os conceitos, mas principalmente saiba aplicá-los, traçamos todo esse panorama através de exercícios práticos.

Questões para revisão

1. Dada a amostra:

   3 - 7 - 10 - 6 - 8 - 6 - 8 - 4 - 5 - 7 - 6 - 10 - 9 - 5 - 6 - 3, responda qual resultado aconteceu com a maior frequência.

   ( ) 4 ( ) 5 ( ) 6

   ( ) 7 ( ) 8

2. Dada a distribuição de frequências, na tabela a seguir,

Idades	Frequência (f)
19 ⊢ 21	8
21 ⊢ 23	12
23 ⊢ 25	15
25 ⊢ 27	13
27 ⊢ 29	7
29 ⊢ 31	5

   responda, assinalando entre os valores da relação abaixo, qual a frequência acumulada total.

   ( ) 31 ( ) 55 ( ) 60

   ( ) 12 ( ) 20

3. Dada a distribuição de frequências na tabela a seguir,

Idades	Frequência (f)
0 ⊢ 2	2
2 ⊢ 4	5
4 ⊢ 6	18
6 ⊢ 8	10
8 ⊢ 10	5

responda, marcando entre as opções da relação que segue, qual o limite superior da quarta classe.

( ) 8 ( ) 6 ( ) 10

( ) 40 ( ) 4

4. Assinale a opção correta. Na distribuição de frequências da questão 3, qual a amplitude de cada classe ou intervalo?

( ) 10 ( ) 40 ( ) 2

( ) 1 ( ) 8

5. Marque a resposta correta. Na distribuição de frequências da questão 3, qual o ponto médio da quinta classe ou intervalo?

( ) 40 ( ) 9 ( ) 8

( ) 5 ( ) 10

6. Observe o quadro a seguir e assinale (V) para verdadeiro e (F) para falso.

	Variável 1	Variável 2	Variável 3	Variável 4	Variável 5
Empregado	Estado civil	N° de dependentes	Altura (em centímetros)	Naturalidade	Sexo
1	S	0	168	RJ	M
2	S	1	174	PR	M
3	C	4	170	SP	F
4	S	0	169	SP	M
5	C	3	178	SP	M
6	C	1	189	RJ	M
7	C	2	170	RS	F
8	S	0	174	SC	F
9	S	0	169	SP	F
10	C	3	177	PR	M

( ) As variáveis 1 e 4 são qualitativas nominais.

( ) As variáveis 4 e 5 são qualitativas ordinais.

( ) As variáveis 2 e 3 são quantitativas discretas.

( ) As variáveis 2 e 3 são quantitativas contínuas.

( ) As variáveis 1 e 5 são qualitativas nominais.

7. Qual a diferença entre variável qualitativa e variável quantitativa?

8. O que é uma distribuição de frequência?

capítulo 4

Medidas de tendência central e de posição

Conteúdos do capítulo

- Como utilizar medidas de posição para resumir dados
- A média aritmética simples para dados não agrupados
- A média aritmética ponderada para dados agrupados
- O valor que ocupa a posição central – mediana
- Valor dos resultados por frequência – moda

Após o estudo deste capítulo, você será capaz de:

1. realizar cálculos com as principais medidas de posição;
2. identificar em que situações as medidas de posição se aplicam.

Medidas de posição

Estudamos, nas unidades anteriores, a sintetização dos dados resultantes de uma pesquisa sob a forma de tabelas, gráficos e de distribuição de frequências que nos permitiam descrever o padrão de variação de um determinado fenômeno estatístico.

Agora, vamos aprender como resumir ainda mais esses dados, apresentando um ou mais valores que sejam representativos da série estudada. São as chamadas *medidas de posição* (ou *medidas de tendência central*), uma vez que representam os fenômenos pelos seus valores médios, em torno dos quais tendem a concentrar-se os dados.

Tais medidas descrevem, de alguma forma, o meio (ou o centro) dos dados.

As medidas de posição (ou de tendência central) mais utilizadas são a média aritmética, a mediana e a moda. As menos utilizadas são a média geométrica, a média harmônica, a quadrática, a cúbica e a biquadrática.[1]

Média aritmética simples

A média aritmética simples, ou simplesmente média, nada mais é que a soma dos resultados obtidos dividida pela quantidade de resultados.

> Vamos estudar inicialmente a média aritmética para dados não agrupados.

Vamos representar a primeira observação por x_1, a segunda observação por x_2 (e assim por diante) e a média aritmética por $\overline{X}$.

Então, se tivermos n observações e desejarmos determinar o valor da média aritmética das mesmas, utilizaremos a fórmula:

$$\overline{X} = \frac{x_1 + x_2 + x_3 + x_4 + \ldots x_n}{n}$$

De forma geral, representamos a média aritmética de dados não agrupados por:

$$\overline{X} = \frac{\Sigma X_i}{n} \text{, em que i varia de 1 até n}$$

ou simplesmente representados por:

$$\overline{X} = \frac{\Sigma X}{n}$$

Exercícios Resolvidos

Agora, vamos aplicar esses conceitos em atividades práticas para que você os fixe bem.

1. Determine a média aritmética dos valores: 5, 8, 10, 12, 15.

Verifique que, como são cinco (5) valores, somaremos os cinco e os dividiremos por cinco.

$$\overline{X} = \frac{\Sigma X}{n} = \frac{5 + 8 + 10 + 12 + 15}{5}$$

Resposta: Portanto, a média aritmética é 10.

2. Em uma empresa de componentes eletrônicos, a exportação nos últimos quatro anos, em milhares de dólares, foi: US$ 800,00, US$ 880,00, US$ 760,00 e US$ 984,00. Determine a média de exportações dessa empresa nesses quatro anos.

Agora, temos quatro (4) valores. Para calcularmos a média deles, somamos os quatro e os dividimos por quatro.

$$\overline{X} = \frac{800 + 880 + 760 + 984}{4}$$

$$\overline{X} = \frac{3424}{4}$$

$\overline{X} = 856{,}00$ milhares de dólares.

Resposta: Logo, a média de exportações da empresa nesses quatro anos foi de 856,00 milhares de dólares.

3. Oito alunos fizeram um teste e obtiveram os seguintes resultados:

9 - 6 - 5 - 8 - 7 - 9 - 4 - 8

Qual a média desses resultados?

$$\overline{X} = \frac{9 + 6 + 5 + 8 + 7 + 9 + 4 + 8}{8}$$

$$\overline{X} = \frac{56}{8}$$

$$\overline{X} = 7$$

Resposta: Portanto, a média dos resultados foi 7.

Média aritmética ponderada

Estudaremos agora a média aritmética para dados agrupados.

Quando os dados estão agrupados numa distribuição de frequências, usamos a média aritmética dos valores $x_1, x_2, x_3, ..., x_n$, ponderados pelas respectivas frequências absolutas $f_1, f_2, f_3, ..., f_n$. Isso significa dizer que cada grandeza envolvida no cálculo da média tem diferente importância ou aconteceu um número diferente de vezes durante a coleta de dados.

Para isso, usamos a fórmula:

$$\overline{X} = \frac{\Sigma (X_i \cdot f_i)}{N}$$, em que i varia de 1 até n

Observe que $N = \Sigma f_i$.

Exercícios Resolvidos

Vamos, então, resolver alguns exercícios com dados agrupados, para você construir a compreensão de seu conhecimento sobre média aritmética ponderada.

1. Calcule a média das idades representadas na distribuição de frequências da Tabela 18.

Tabela 18 – Idades de um grupo de pessoas

Idade	Frequência (f)
4	4
5	6
6	6
7	4

em que i varia de 1 até n 55

$$\overline{X} = \frac{4 \cdot 4 + 5 \cdot 6 + 6 \cdot 6 + 7 \cdot 4}{20}$$

$$\overline{X} = \frac{16 + 30 + 36 + 28}{20}$$

$$\overline{X} = \frac{110}{20} = 5{,}5$$

Resposta: A média das idades desse grupo de pessoas é de 5,5 anos.

2. Calcule a média das idades representadas na distribuição de frequências da Tabela 19.

Tabela 19 – Idades de um grupo de pessoas

Idades	Frequência (f)
18 ├── 21	9
21 ├── 24	12
24 ├── 27	12
27 ├── 30	17
30 ├── 33	16
33 ├── 36	14
36 ├── 39	11
39 ├──┤ 42	9

Nesse caso, em que a variável estudada (idade) está representada em intervalos (ou classes), os valores de X_i na fórmula da média aritmética são representados pelos pontos médios (Pm_i) desses intervalos. Assim, temos:

$$\overline{X} = \frac{\Sigma (Pm_i \cdot f_i)}{N}$$

$$\overline{X} = \frac{19{,}5 \cdot 9 + 22{,}5 \cdot 12 + 25{,}5 \cdot 12 + 28{,}5 \cdot 17 + 31{,}5 \cdot 16 + 34{,}5 \cdot 14 + 37{,}5 \cdot 11 + 40{,}5 \cdot 9}{9 + 12 + 12 + 17 + 16 + 14 + 11 + 9}$$

$$\overline{X} = \frac{175{,}5 + 270 + 306 + 484{,}5 + 504 + 483 + 412{,}5 + 364{,}5}{100}$$

$$\overline{X} = \frac{3.000}{100}$$

$\overline{X}$ = 30 anos

Resposta: A média das idades desse grupo de pessoas é de 30 anos.

3. Calcule a média da turma que realizou um teste de Estatística e obteve os resultados representados na tabela 20.

Tabela 20 – Notas obtidas em um teste de Estatística

Idades	Frequência (f)
0 ⊢ 1	4
1 ⊢ 2	5
2 ⊢ 3	7
3 ⊢ 4	8
4 ⊢ 5	9
5 ⊢ 6	14
6 ⊢ 7	8
7 ⊢ 8	6
8 ⊢ 9	5
9 ⊢⊣ 10	4
Σ	70

Novamente os resultados da variável estudada estão representados em intervalos (ou classes); então, de novo utilizaremos os pontos médios dos intervalos, para determinar a média aritmética procurada.

$$\overline{X} = \frac{0,5 \cdot 4 + 1,5 \cdot 5 + 2,5 \cdot 7 + 3,5 \cdot 8 + 4,5 \cdot 9 + 5,5 \cdot 14 + 6,5 \cdot 8 + 7,5 \cdot 6 + 8,5 \cdot 5 + 9,5 \cdot 4}{4 + 5 + 7 + 8 + 9 + 14 + 8 + 6 + 5 + 4}$$

$$\overline{X} = \frac{2 + 7,5 + 17,5 + 28 + 40,5 + 77 + 52 + 45 + 42,5 + 38}{70}$$

$$\overline{X} = \frac{350}{70}$$

$$\overline{X} = 5$$

Resposta: Já esse grupo de Estatística obteve a média 5.

IMPORTANTE! É comum representar-se por $\overline{X}$ a média aritmética de uma amostra e por μ* a média aritmética de uma população. Nós só utilizaremos, para os dois casos, a representação $\overline{X}$.

Mediana

A mediana de um conjunto de dados é o valor que ocupa a posição central desses dados, desde que estejam colocados em ordem crescente ou decrescente, ou seja, em um rol.

Nós representaremos a mediana de uma amostra ou de uma população por Md.

É necessário, entretanto, observar que a quantidade de dados pode ser par ou ímpar. Sendo ímpar, o valor da mediana é o valor que está no centro da série; sendo par, a mediana será a média aritmética dos dois valores que estão no centro da série.

Vamos analisar alguns exemplos para o caso da variável discreta.

Calcule a mediana dos dados:

5 - 8 - 4 - 6 - 7 - 3 - 4

O primeiro passo é colocá-los em ordem crescente:

3 - 4 - 4 - 5 - 6 - 7 - 8

* μ é a letra grega minúscula que lemos "mu" e que corresponde à letra *m* do nosso alfabeto.

O SEGUNDO PASSO é verificar se a quantidade de dados é par ou ímpar. Nesse caso, é ímpar; então, a mediana é valor central da série, ou seja:

Md = 5

Agora,

Calcule a mediana dos dados:

8 - 0 - 7 - 4 - 7 - 10 - 6 - 5

Colocando os dados em ordem crescente, temos:

0 - 4 - 5 - 6 - 7 - 7 - 8 - 10

Como agora a quantidade de dados é par, a mediana é a média aritmética dos dois valores centrais da série, ou seja:

$$Md = \frac{6 + 7}{2}$$

Md = 6,5

Vamos analisar outra situação.

Em uma pesquisa em que desejamos conhecer a altura dos 50 alunos de uma classe, os resultados obtidos nas medições, em centímetros, e colocados em ordem crescente foram os representados na Tabela 21.

Tabela 21 – Altura dos alunos de uma classe

Alturas (cm)	Frequência (f)	Frequência acumulada (f_a)
160	2	2
161	1	3
162	1	4
163	1	5
164	2	7
165	5	12
166	2	14
167	4	18
168	4	22
169	1	23
170	7	30
171	1	31
172	2	33
173	1	34
174	4	38
175	2	40
176	2	42
177	3	45
178	3	48
179	1	49
180	1	50
Σ	50	

Como calcular a altura mediana desses alunos?

O primeiro passo é determinar se n é par ou ímpar. No caso, n = 50, ou seja, é par. Então, a mediana será obtida por meio dos dois elementos centrais da série.

O segundo passo é identificar, por intermédio da frequência acumulada, onde se encontram esses dois elementos, ou seja, o 25^0 e o 26^0 elementos. Olhando para a tabela anterior, você verifica que tanto o 25^0 quanto o 26^0 elementos têm 170 centímetros. Então, a mediana é:

$$Md = \frac{170 + 170}{2}$$

Md = 170 centímetros

Observe que pode acontecer de, em 50 alunos, nenhum ter uma determinada medida (por exemplo, 168 cm). Nesse caso, f = 0 e os cálculos são efetuados como no caso anterior.

E, COMO FARÍAMOS ESSE CÁLCULO, SE A VARIÁVEL PESQUISADA FOSSE CONTÍNUA E COM OS DADOS AGRUPADOS EM CLASSES (OU INTERVALOS)? Vejamos, no exemplo seguinte, a tabela anterior representada em sete classes (ou intervalos), conforme a tabela 22.

Tabela 22 – Altura dos alunos de uma classe

Alturas (cm)	Frequência (f)	Frequência acumulada (f_a)
160 ⊢ 163	4	4
163 ⊢ 166	8	12
166 ⊢ 169	10	22
169 ⊢ 172	9	31
172 ⊢ 175	7	38
175 ⊢ 178	7	45
178 ⊢⊣ 181	5	50
Σ	50	

O PRIMEIRO PASSO consiste em determinar o valor de n/2. Como a variável é contínua, não se preocupe se n é um número par ou ímpar. No caso, como n = 50, n / 2 é igual a 25.

O SEGUNDO PASSO é identificar em que classe (ou intervalo) está contida a mediana. Observando-se a tabela anterior, verificamos que o elemento de ordem 25 encontra-se no quarto intervalo, ou classe (cujos valores estão compreendidos entre 169 e 172 centímetros).

O PRÓXIMO PASSO é o cálculo da mediana, utilizando-se a fórmula:

$$Md = Li + \frac{(n/2 - \Sigma f_{ant})}{f_{Md}} \cdot A$$

Em que:

> Li = limite inferior da classe que contém a mediana.
> n = tamanho da amostra ou da população que estamos pesquisando.
> Σf_{ant} = soma das frequências anteriores à classe que contém a mediana.
> A = amplitude da classe que contém a mediana.
> f_{Md} = frequência da classe que contém a mediana.

Vamos, então, ao cálculo da mediana.

$$Md = 169 + \frac{(25 - 22)}{9} \cdot 3$$

$$Md = 169 + \frac{9}{9}$$

Md = 170 CENTÍMETROS

Exercício Resolvido

Perguntamos a idade a um grupo de 100 pessoas, com o objetivo de calcular a idade mediana desse grupo, e representamos os resultados obtidos na Tabela 23, a seguir, de distribuição de frequências.

Tabela 23 – Idades de um grupo de pessoas

Idades	Frequência (f)	Frequência acumulada (f_a)
18 ⊢ 21	9	9
21 ⊢ 24	12	21
24 ⊢ 27	12	33
27 ⊢ 30	17	50
30 ⊢ 33	16	66
33 ⊢ 36	14	80
36 ⊢ 39	11	91
39 ⊢⊣ 42	9	100
Σ	100	

Nesse caso, o primeiro passo é determinar n/2 = 100/2 = 50.

O segundo passo é verificar a classe em que se encontra o 50^0 elemento da série. No caso, está na quarta classe, cujas idades vão de 27 a 30 anos. Não se preocupe se o intervalo é aberto ou fechado.

O terceiro passo é utilizar a fórmula para o cálculo da mediana.

$$Md = Li + \frac{(n/2 - \Sigma f_{ant}) \cdot A}{f_{Md}}$$

$$Md = 27 + \frac{(50 - 33)}{17} \cdot 3$$

$$Md = 27 + \frac{51}{17}$$

RESPOSTA: Md = 30 anos

Moda

Como você percebe, andando na rua, que uma roupa está na moda?

Você percebe isso porque vê muitas pessoas com aquela roupa. Em outras palavras, é uma roupa que aparece com muita frequência.

Evidentemente, esse é um exemplo bem simplista.

> Na estatística, nós definiremos *moda* como sendo o valor dos resultados de uma pesquisa que acontece com a maior frequência e a representaremos por *Mo*.

Nas distribuições sem o agrupamento em intervalos (ou classes), a simples observação da coluna das frequências nos permite saber qual é o elemento da série que apresenta a maior frequência e que, portanto, é o valor que está na moda; nós o chamamos de VALOR MODAL.

Assim, para a distribuição

X_i	f_i
5	3
6	5
7	11
8	10
9	5

a moda é o 7. Representa-se por Mo = 7.

Para os dados agrupados em intervalos (ou classes), a determinação do valor modal também é simples, porém mais trabalhoso; trata-se de um processo semelhante ao utilizado para o cálculo da mediana.

Na realidade, existe mais de um método para a determinação da moda. O mais simples e, consequentemente, mais impreciso, é considerar como moda o ponto médio da classe que contém a moda, também denominada de *classe modal*; esse valor chama-se MODA BRUTA.

Podemos, ainda, determinar a moda mediante o uso de fórmulas ou de gráficos. Para tal, temos o método de King e o método de Czuber*. Nós utilizaremos o método de King.

* Emanuel Czuber, nascido em Praga (1851-1925).

Determinaremos a moda da distribuição de frequências representada na Tabela 24.

Tabela 24 – Idades de um grupo de pessoas

Idades	Frequência (f)	Frequência acumulada (f_a)
18 ├── 21	9	9
21 ├── 24	12	21
24 ├── 27	12	33
27 ├── 30	17	50
30 ├── 33	16	66
33 ├── 36	14	80
36 ├── 39	11	91
39 ├──┤ 42	9	100
Σ	100	

O PRIMEIRO PASSO consiste em identificar em que classe se encontra a moda, ou seja, qual é a classe modal. No caso, é a quarta classe (idades de 27 a 30 anos), pois é a que apresenta a maior frequência de ocorrência (f = 17).

O SEGUNDO PASSO é a determinação do valor da moda por meio da fórmula:

$$Mo = Li + \frac{f_{post} \cdot A}{f_{ant} + f_{post}}$$

Em que:

Li = limite inferior da classe que contém a moda.

f_{post} = frequência da classe posterior à classe que contém a moda.

f_{ant} = frequência da classe anterior à classe que contém a moda.

A = amplitude da classe que contém a moda.

Vamos, agora, substituir os elementos dessa fórmula pelos valores do nosso problema; assim, temos:

$$Mo = 27 + \frac{16}{12 + 16} \cdot 3$$

$$Mo = 27 + \frac{48}{28}$$

Mo = 28,7143 ANOS.

IMPORTANTE! Sendo conhecidas a média aritmética e a mediana de uma série, é possível obtermos o valor da moda pela aplicação da fórmula de Pearson*, em que: $Mo = 3 \cdot Md - 2 \cdot \bar{X}$

Essa fórmula nos dá o valor aproximado da moda e só deve ser utilizada quando a distribuição apresentar razoável simetria em relação à média.

Vamos, agora, determinar a moda da distribuição de frequências representada na Tabela 25.

Tabela 25 – Alturas de um grupo de pessoas

Alturas (cm)	Frequência (f_i)	Frequência acumulada (f_a)
160 ⊢ 163	4	4
163 ⊢ 166	8	12
166 ⊢ 169	10	22
169 ⊢ 172	9	31
172 ⊢ 175	7	38
175 ⊢ 178	7	45
178 ⊢ 181	5	50
Σ	50	

O PRIMEIRO PASSO é a constatação, visual, da classe que contém a moda. No caso, a classe modal é a terceira, cuja frequência de ocorrência é igual a 10.

Então, a moda é:

$$Mo = 166 + \frac{9}{8+9} \cdot 3$$

Mo = 166 + 1,588

Mo = 167,588 CENTÍMETROS.

* Karl Pearson, nascido em Londres (1857-1936), foi um grande contribuidor para o desenvolvimento da Estatística como uma disciplina científica séria e independente.

> IMPORTANTE! Pode ocorrer que uma série apresente mais de uma moda. A plurimodalidade pode ocorrer em função de:
> dados pertencentes a populações diferentes;
> insuficiência de dados para análise;
> quantidade inadequada de classes (ou intervalos).

Síntese

A significativa importância das medidas de posição está no fato de permitirem reduzir as variáveis de uma série estudada a um ou mais valores. Isso é possível, pois, como você viu, elas nos permitem estabelecer os valores médios. Para realizar tal processo, precisamos trabalhar com a média aritmética simples, no caso de dados não agrupados, e com a média aritmética ponderada, quando utilizamos dados agrupados. Outros conceitos que precisamos dominar e saber calcular no processo estatístico são os de mediana e de moda, razão pela qual utilizamos vários exemplos para facilitar a elaboração de sua compreensão dos cálculos e usos desses saberes.

Questões para revisão

1. Assinale, nos parênteses, a opção correta. Dados os valores a seguir,

   9 - 6 - 5 - 4 - 8 - 9 - 10 - 4 - 7 - 8 - 5 - 6 - 10,

   determine a média aritmética deles.

   ( ) 9 ( ) 7,5 ( ) 6,5

   ( ) 8,5 ( ) 7

2. Em uma pesquisa realizada em uma empresa quanto aos salários médios de seus funcionários, verificou-se o seguinte resultado:

Salários (R$)	Frequência (f)
240,00 ⊢ 480,00	15
480,00 ⊢ 720,00	22
720,00 ⊢ 960,00	30
960,00 ⊢ 1.200,00	18
1.200,00 ⊢⊣ 1.440,00	15

Baseado nesses resultados, determine o salário médio desses funcionários e marque a resposta correta nas opções abaixo.

( ) R$ 720,00; ( ) R$ 840,00;

( ) R$ 830,40; ( ) R$ 1.000,00.

( ) R$ 954,60;

3. Dados os valores a seguir,

   9 - 6 - 5 - 4 - 8 - 9 - 10 - 4 - 7 - 8 - 5 - 6 - 10,

   determine a mediana dos mesmos e assinale, na relação abaixo, a resposta correta.

   ( ) 7 ( ) 9,5 ( ) 6

   ( ) 10 ( ) 7,5

4. Assinale a resposta correta, após determinar, pela fórmula, a mediana da distribuição apresentada no exercício 2.

   ( ) R$ 960,00; ( ) R$ 840,00;

   ( ) R$ 720,00; ( ) R$ 828,00.

   ( ) R$ 824,00;

5. Marque a opção correspondente, após determinar, pela fórmula, a moda da distribuição apresentada no exercício 2.

   ( ) R$ 960,00; ( ) R$ 840,00;

   ( ) R$ 720,00; ( ) R$ 828,00.

   ( ) R$ 824,00;

6. As exportações de determinado porto brasileiro registraram o seguinte movimento, em bilhões de reais, mês a mês, durante um ano:

Mês	R$
Janeiro	3,0
Fevereiro	2,4
Março	2,8
Abril	3,1
Maio	2,7
Junho	3,2
Julho	2,6
Agosto	2,5
Setembro	3,4
Outubro	3,4
Novembro	3,3
Dezembro	3,6
Σ	36,0

Qual foi a média mensal de exportações, em bilhões de reais? Marque a alternativa correta.

( ) 3,0 ( ) 3,2 ( ) 2,6

( ) 2,9 ( ) 36,0

7. O comércio varejista de calçados registrou as seguintes encomendas, em milhares de pares, em uma determinada indústria calçadista, ao longo de um semestre:

Mês	Pares encomendados
1° mês	48
2° mês	60
3° mês	42
4° mês	50
5° mês	64
6° mês	66
Σ	330

Qual a média mensal de encomendas, em milhares de pares de calçados? Assinale a média correta.

( ) 42 ( ) 33 ( ) 330

( ) 50 ( ) 55

8. O Departamento de Marketing e Propaganda de uma multinacional investiu, durante um ano, os seguintes valores em campanhas publicitárias, em milhares de reais:

Mês	R$
Janeiro	20
Fevereiro	12
Março	8
Abril	24
Maio	0
Junho	16
Julho	18
Agosto	20
Setembro	12
Outubro	15
Novembro	13
Dezembro	22
Σ	180

Qual foi a média mensal de investimento, em milhares de reais? Assinale a alternativa correta.

( ) 18 ( ) 12 ( ) 20

( ) 15 ( ) 17

9. Determinada máquina produziu durante oito dias as seguintes quantidades de peças (em unidades):

Dia	Unidades
1°	140
2°	136
3°	142
4°	128
5°	154
6°	140
7°	158
8°	170
Σ	1.168

Qual a média diária de produção dessa máquina? Assinale a resposta correta.

( ) 128 ( ) 141 ( ) 146

( ) 154 ( ) 140

10. Uma secretária fez uma pesquisa de preço para a aquisição de uma impressora da marca X e obteve os seguintes valores, em reais:

Loja	Valor (R$)
A	484,00
B	399,00
C	505,00
D	383,00
E	429,00
Σ	2.200,00

Qual o preço médio dessa pesquisa? Assinale a resposta correta.

( ) 505,00 ( ) 440,00 ( ) 399,00

( ) 484,00 ( ) 444,00

11. Para que servem as medidas de posição?

12. Descreva a diferença entre média aritmética simples e média aritmética ponderada.

capítulo 5

Medidas de dispersão

Conteúdos do capítulo

- Conceito e uso das medidas de dispersão.
- Os fenômenos de amplitude total, amplitude semi-interquartílica, desvio médio, variância e desvio padrão.

Após o estudo deste capítulo, você será capaz de:

1. realizar cálculos com as principais medidas de dispersão;
2. identificar as medidas de dispersão.

Medidas de dispersão

As medidas de dispersão (ou de afastamento) são medidas estatísticas utilizadas para verificar o quanto os valores encontrados em uma pesquisa estão dispersos ou afastados em relação à média ou em relação à mediana.

São medidas que servem para verificar com que confiança as medidas de tendência central resumem as informações fornecidas pelos dados obtidos em uma pesquisa.

Imaginemos, por exemplo, duas pessoas que tenham se submetido a um teste. Suponhamos duas situações diferentes:

a) as duas pessoas tiraram nota igual a 6,0;

b) as duas pessoas tiraram, respectivamente, nota 2,0 e nota 10,0.

Nos dois casos, as duas pessoas obtiveram média igual a 6,0. No entanto, no caso “a” elas se concentraram sobre a média e no caso “b” dispersaram-se (afastaram-se) da média.

Isso significa dizer que a média é muito mais representativa no caso “a” do que no caso “b”. Mostra, também, que no caso “a” existe uma homogeneidade nos conhecimentos adquiridos pelas pessoas, enquanto no caso “b” existe uma heterogeneidade.

Para avaliarmos o grau de variabilidade (ou dispersão, ou afastamento) dos valores de um conjunto de números, utilizamos as medidas de dispersão absoluta, que nos permitem obter um conhecimento mais completo e detalhado do fenômeno em questão. São elas:

a) amplitude total ou intervalo total;

b) amplitude semi-interquartílica, ou intervalo semi-interquartílico, ou desvio quartil;

c) desvio médio;

d) variância;

e) desvio padrão.

Amplitude total (ou intervalo total)

A AMPLITUDE TOTAL (ou intervalo total) é a diferença entre o maior e o menor valor de uma série de dados. Por exemplo, no conjunto de números

4, 6, 8, 9, 12, 17, 20, a amplitude total será A = 20 – 4 = 16.

Para o caso dos dados estarem agrupados em classes (ou intervalos), o cálculo da amplitude total pode ser feito de duas formas:

a) pelo PONTO MÉDIO DAS CLASSES; nesse caso, a amplitude total é igual ao ponto médio da última classe, menos o ponto médio da primeira classe;

b) pelos LIMITES DAS CLASSES; nesse caso, a amplitude total é igual ao limite superior da última classe, menos o limite inferior da primeira classe.

Por exemplo, veja a distribuição de frequências representada na tabela 26.

Tabela 26 – Idades de um grupo de pessoas

Idades	Frequência (f_i)
18 ⊢ 21	9
21 ⊢ 24	12
24 ⊢ 27	12
27 ⊢ 30	17
30 ⊢ 33	16
33 ⊢ 36	14
36 ⊢ 39	11
39 ⊢⊣ 42	9

A amplitude total é:

a) pelo ponto médio das classes,

A = 40,5 – 19,5 = 21

b) pelos limites das classes,

A = 42 – 18 = 24

A AMPLITUDE TOTAL (ou intervalo total), embora seja a medida de dispersão mais simples, apresenta algumas restrições quanto ao seu uso, porque é muito instável. Essa instabilidade ocorre em virtude da amplitude total levar em consideração apenas os valores extremos da série, não sendo afetada pela dispersão dos seus valores internos.

Amplitude semi-interquartílica

Antes de explicarmos o que é AMPLITUDE SEMI-INTERQUARTÍLICA, é necessário esclarecer que temos medidas de posição semelhantes – na sua concepção – à mediana, apesar de não serem medidas de tendência central. São elas os quartis, os decis e os percentis.

Os QUARTIS permitem dividir a distribuição em quatro partes iguais, ou seja, com o mesmo número de elementos. Assim, os DECIS dividem a distribuição em dez partes iguais e os PERCENTIS em cem partes iguais. Em relação aos quartis, representamos a seguir um conjunto de dados em que seus elementos foram divididos em quatro partes iguais:

0% 25% 50% 75% 100%

Q_1 Q_2 Q_3

Em que:

Q_1 = primeiro quartil (Q_1 é a mediana de todos os valores inferiores à mediana do conjunto de dados).

Q_2 = segundo quartil (coincide com a mediana).

Q_3 = terceiro quartil (Q_3 é a mediana de todos os valores superiores à mediana do conjunto de dados).

Para o cálculo de Q_1, Q_2 e Q_3, usamos, quanto às fórmulas, a seguinte sequência:

$$Q_1 = Li + \frac{(n/4 - \Sigma f_{ant}) \cdot A}{f_{Q_1}}$$

$$Q_2 = Li + \frac{(n/2 - \Sigma f_{ant}) \cdot A}{f_{Q_2}}$$

$$Q_3 = Li + \frac{(3n/4 - \Sigma f_{ant}) \cdot A}{f_{Q_3}}$$

A AMPLITUDE SEMI-INTERQUARTÍLICA, também denominada de *intervalo semi-interquartílico* ou de *desvio quartil*, é utilizada para verificar a dispersão em relação à mediana. Seu cálculo leva em consideração o primeiro e o terceiro quartis e é efetuado pela fórmula:

$$D_q = \frac{Q_3 - Q_1}{2}$$

Em que:

D_q = desvio quartil.

Q_3 = terceiro quartil.

Q_1 = primeiro quartil.

É importante observarmos que no desvio quartil, no intervalo definido pelos limites ($Md - D_q$) e ($Md + D_q$), encontram-se 50% dos elementos, aproximadamente. Se a distribuição for simétrica, essa porcentagem será exata.

Outra grande importância do desvio quartil é o fato de ser uma medida que não é afetada pelos valores extremos da série, sendo, portanto, útil quando esses valores extremos não forem representativos.

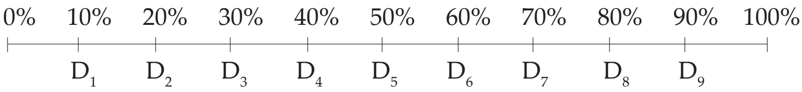
Como exemplo de aplicação, vamos supor que uma turma de alunos fez um teste de Estatística e um teste de Matemática, nos quais obteve os seguintes resultados:

a) Estatística: $Md = 6$, $Q_1 = 4{,}5$, $Q_3 = 7{,}5$

b) Matemática: $Md = 6$, $Q_1 = 3$, $Q_3 = 7{,}5$

Observamos que nos dois testes a mediana foi igual a 6,0. No entanto, as notas de Estatística apresentam uma dispersão menor que as notas de Matemática.

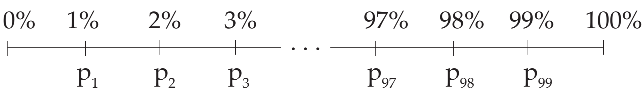
Quanto aos decis, representamos, a seguir, um conjunto de dados em que seus elementos foram divididos em dez partes iguais.



Em que:

D_1 = primeiro decil.
D_2 = segundo decil.
⋮
D_9 = nono decil.

Finalmente, quanto aos percentis, representamos a seguir um conjunto de dados em que seus elementos foram divididos em 100 partes iguais.

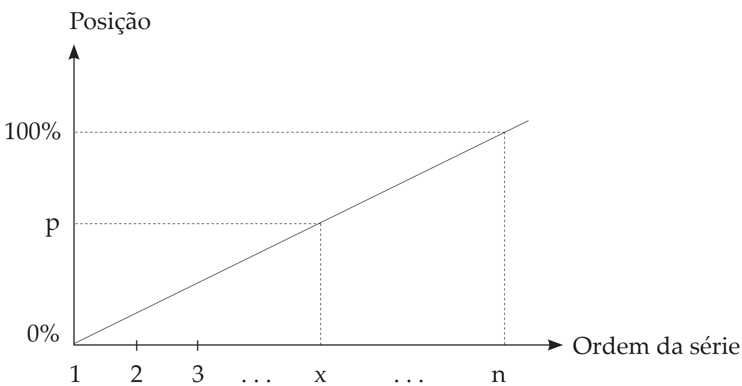


Em que:

p_1 = primeiro percentil.
p_2 = segundo percentil.
p_3 = terceiro percentil.
⋮
p_{98} = nonagésimo oitavo percentil.
p_{99} = nonagésimo nono percentil.

De acordo com Lapponi[1], após ordenar a série de n observações em ordem crescente de valores, definimos como 0% a posição da observação de ordem 1 e como 100% a posição da observação de ordem n.

Gráfico 6 – Ordenamento de uma série de n observações



Como calcular o percentil correspondente ao valor de ordem *x*?

Esse cálculo é efetuado a partir da expressão:

$$p = \frac{x - 1}{n - 1} \cdot 100\ \%$$

Em que:

n = número total de observações da série.
x = ordem de uma determinada observação.
p = percentil (em percentual) da observação.

Exercício Resolvido

1. Vamos supor uma série de 20 observações registradas no quadro a seguir e calcular: qual o valor do percentil cuja ordem é x = 1? e, qual o valor do percentil cuja ordem é x = 8?

8	27	18	22	24	10	11	9	17	16
29	26	20	13	28	30	12	21	15	25

Vamos colocar essa série ordenada de forma crescente e associar a cada valor da série um número de ordem de 1 a 20.

Observação da série	8	9	10	11	12	13	15	16	17	18
Número de ordem	1	2	3	4	5	6	7	8	9	10

Observação da série	20	21	22	24	25	26	27	28	29	30
Número de ordem	11	12	13	14	15	16	17	18	19	20

Qual o valor do percentil cuja ordem é x = 1?

$$p = \frac{1 - 1}{20 - 1} \cdot 100\ \%$$

p = 0%

E qual o valor do percentil cuja ordem é x = 8?

$$p = \frac{8 - 1}{20 - 1} \cdot 100\ \%$$

$$p = \frac{7}{19} \cdot 100\ \%$$

$$p = 36{,}8\%$$

Desvio médio

Quando desejamos analisar a dispersão (ou afastamento) dos valores de uma série em relação à média, é conveniente analisar essa dispersão de cada um dos valores, sem exceção. Assim, chamaremos *Dm* de *desvio médio* e o calcularemos pela fórmula:

$$Dm = \frac{\Sigma\ |X_i - \overline{X}| \cdot f_i}{N}$$

Em que $| X_i - \overline{X} |$ é o módulo de cada desvio em relação à média, i varia de 1 até n e N é igual a Σf.

> IMPORTANTE! Estamos trabalhando com o módulo do desvio porque, de acordo com a propriedade de média aritmética, a soma de todos os desvios em relação à média é sempre igual a zero. Trabalhando com o módulo (que é sempre positivo), eliminamos esse problema.

Pela fórmula acima, deduzimos que o desvio médio nada mais é do que a média aritmética dos desvios dos valores em relação à média, considerados em valor absoluto.

Como exemplo de aplicação, calculemos o desvio médio do seguinte conjunto de números:

4, 6, 8, 9, 10, 11

Inicialmente, devemos calcular a média aritmética dos valores dados:

$$\overline{X} = \frac{4 + 6 + 8 + 9 + 10 + 11}{6}$$

$$\overline{X} = 8$$

Temos, então:

X_i	$X_i - \overline{X}$	$\mid X_i - \overline{X} \mid$
4	4 – 8 = – 4	4
6	6 – 8 = – 2	2
8	8 – 8 = 0	0
9	9 – 8 = 1	1
10	10 – 8 = 2	2
11	11 – 8 = 3	3
Σ	0	12

Portanto, o desvio médio é:

$$Dm = \frac{12}{6}$$

$$Dm = 2$$

Observe que, nesse exemplo, todos os valores ocorreram uma única vez. Por isso, em todos eles a frequência f foi considerada igual a 1.

Para dados agrupados em classes ou intervalos, substitui-se X_i, na fórmula, pelo ponto médio de cada classe.

Vamos agora calcular o desvio médio da distribuição de frequências da tabela 26.

O primeiro passo consiste em calcular a média das idades. Observe que tal média já foi calculada na página 55, quando verificamos que $\overline{X} = 30$. Vamos, então, verificar o quanto cada idade está afastada da média ($X_i - \overline{X}$).

O segundo passo é montar uma tabela com cálculo do desvio de cada idade em relação à média encontrada. Não esquecer que, para o cálculo do desvio médio, é necessário multiplicar o módulo de cada desvio encontrado pela respectiva frequência (f_i).

Tabela 27 – Desvio médio de cada idade em relação à média

Idades	Frequência (f)	$X_i - \overline{X}$	$\mid X_i - \overline{X} \mid . f_i$
18 ⊢ 21	9	19,5 – 30 = – 10,5	10,5 · 9 = 94,5
21 ⊢ 24	12	22,5 – 30 = – 7,5	7,5 · 12 = 90
24 ⊢ 27	12	25,5 – 30 = – 4,5	4,5 · 12 = 54
27 ⊢ 30	17	28,5 – 30 = – 1,5	1,5 · 17 = 25,5
30 ⊢ 33	16	31,5 – 30 = + 1,5	1,5 · 16 = 24
33 ⊢ 36	14	34,5 – 30 = + 4,5	4,5 · 14 = 63
36 ⊢ 39	11	37,5 – 30 = + 7,5	7,5 · 11 = 82,5
39 ⊢⊣ 42	9	40,5 – 30 = + 10,5	10,5 · 9 = 94,5
Σ	100		528

O desvio médio dessa distribuição de frequências é igual a:

$$Dm = \frac{\Sigma\ |X_i - \overline{X}| \cdot f_i}{N}$$

$$Dm = \frac{528}{100}$$

$$D_M = 5{,}28$$

Variância (S^2)

Já vimos que a soma dos desvios em relação à média é sempre igual a zero. Por isso, consideramos os valores de $X_i - \overline{X}$ em módulo. Outra forma de evitarmos que esse fato interfira nos nossos cálculos é considerar cada desvio elevado ao quadrado, pois sabemos que o quadrado de qualquer número real é sempre positivo.

À média aritmética dos quadrados dos desvios damos o nome de *variância*.

Sua fórmula, quando se trata de população, é:

$$S^2 = \frac{\Sigma\ (X_i - \overline{X})^2 \cdot f_i}{N}$$

Para o cálculo da variância de uma amostra, o denominador dessa fórmula deverá ser (N – 1), ou seja:

$$S^2 = \frac{\Sigma\ (X_i - \overline{X})^2 \cdot f_i}{N - 1}$$

Por que (N – 1)?

Caso tomássemos muitas amostras de uma população que tem média μ e calculássemos as médias amostrais $\overline{X}$ de todas essas amostras, verificaríamos que a média dessas médias amostrais ficaria muito próxima a μ.

Entretanto, se calculássemos a variância dessas amostras pela fórmula

$$S^2 = \frac{\Sigma\ (X_i - \overline{X})^2 \cdot f_i}{N}$$

e, então, determinássemos a média dessas variáveis, verificaríamos que essa média é menor que a variância da população. Teoricamente, pode

ser mostrado que podemos compensar isso dividindo a fórmula de S^2 por (N – 1), quando se tratar de uma amostra.[2]

Verifica-se, segundo Toledo e Ovalle[3], entretanto, que para valores grandes de N (N > 30) não há grande diferença entre os resultados proporcionados pela utilização de qualquer dos dois divisores, N ou (N – 1).

Desvio padrão (S)

O cálculo da variância é, na verdade, um passo intermediário para o cálculo do que denominamos de *desvio padrão*.

Para o cálculo de S, vamos extrair a raiz quadrada da variância para compensar o fato de termos elevado ao quadrado os desvios em relação à média.

O DESVIO PADRÃO é a medida de dispersão mais utilizada na prática, considerando, tal qual o desvio médio, os desvios em relação à média. Nós o representamos por S.

Para que o desvio padrão tenha um significado quantitativo, vamos adiantar conceitos que não fazem parte do presente capítulo, mas que você pode estudar quando achar necessário.

Para um conjunto de dados obtidos por amostragem, de uma população a que denominamos de *normal*, verificamos que, quando a amostra é suficientemente grande, o intervalo compreendido entre os valores $(\overline{X} - S)$ e $(\overline{X} + S)$ inclui aproximadamente 68,26% dos resultados obtidos na pesquisa e o intervalo compreendido entre $(\overline{X} - 2 \cdot S)$ e $(\overline{X} + 2 \cdot S)$ inclui aproximadamente 95,44% desses resultados. Se considerarmos o intervalo compreendido entre $(\overline{X} - 3 \cdot S)$ e $(\overline{X} + 3 \cdot S)$, temos cerca de 99,74% dos resultados, ou seja, quase a totalidade das observações efetuadas.

Caso façamos o histograma correspondente, vemos que ele se aproxima de uma curva denominada de *curva de Gauss*, ou *curva em forma de sino*, ou, ainda, *curva normal*.

Para o cálculo do desvio padrão de uma população temos a fórmula:

$$S = \sqrt{\frac{\Sigma (X_i - \overline{X})^2 \cdot f_i}{N}}$$

A fórmula para o cálculo do desvio padrão da amostra é, portanto:

$$S = \sqrt{\frac{\Sigma (X_i - \overline{X})^2 \cdot f_i}{N - 1}}$$

Lembrar que $\sqrt{S^2} = S$

Exercícios Resolvidos

1. Oito empregados de uma Siderúrgica que exercem a mesma função têm os seguintes salários: R$ 2.800,00; R$ 2.650,00; R$ 2.920,00; R$ 2.800,00; R$ 2878,00; R$ 2.682,00; R$ 2.700,00; R$ 2.570,00. Qual seria o padrão desses salários, considerando-os como uma população?

$$\overline{X} = \frac{2800 + 2650 + 2920 + 2800 + 2878 + 2682 + 2700 + 2570}{8}$$

$$\overline{X} = 2.750{,}00$$

$$S^2 =$$

$$= \frac{(2800 - 2750)^2 + (2650 - 2750)^2 + (2920 - 2750)^2 + (2800 - 2750)^2 + (2878 - 2750)^2 + (2682 - 2750)^2 + (2700 - 2750)^2 + (2570 - 2750)^2}{8}$$

$$S^2 = \frac{2500 + 10000 + 28900 + 2500 + 16384 + 4624 + 2500 + 32400}{8}$$

$$S^2 = 12.476{,}00$$

Observar que a frequência foi considerada sempre igual a 1.

Então:

$$S = \sqrt{12476}$$

$$S = 111{,}70$$

2. Considerando-se agora que a tabela 28 é uma amostra das idades de um grupo de pessoas, vamos determinar o desvio padrão da distribuição.

Sabemos que:

$$S = \sqrt{\frac{\Sigma (X_i - \overline{X})^2 \cdot f_i}{N - 1}}$$

Sabemos, também, que $\overline{X} = 30$

Tabela 28 – Quadrado do desvio médio de cada idade em relação à média

Idades	Frequência (f_i)	$\lvert X_i - \overline{X} \rvert$	$(\lvert X_i - \overline{X} \rvert)^2 \cdot f_i$
18 ⊢ 21	9	19,5 – 30 = 10,5	$10,5^2 \cdot 9 = 992,25$
21 ⊢ 24	12	25,5 – 30 = 7,5	$7,5^2 \cdot 12 = 675$
24 ⊢ 27	12	22,5 – 30 = 4,5	$4,5^2 \cdot 12 = 243$
27 ⊢ 30	17	28,5 – 30 = 1,5	$1,5^2 \cdot 17 = 38,25$
30 ⊢ 33	16	31,5 – 30 = 1,5	$1,5^2 \cdot 16 = 36$
33 ⊢ 36	14	34,5 – 30 = 4,5	$4,5^2 \cdot 14 = 283,5$
36 ⊢ 39	11	37,5 – 30 = 7,5	$7,5^2 \cdot 11 = 618,75$
39 ⊢⊣ 42	9	40,5 – 30 = 10,5	$10,5^2 \cdot 9 = 992,25$
Σ	100		3.879

Então, de acordo com os valores obtidos na Tabela 28, o desvio padrão procurado é:

$$S = \sqrt{\frac{3879}{99}}$$

Resposta: S = 6,26

3. Dada a tabela a seguir, em que constam as medidas obtidas em uma população quanto aos seus salários recebidos (em quantidade de salários mínimos), vamos determinar o desvio padrão da distribuição.

Tabela 29 – Salários de uma população observada

Salários mínimos	N° de pessoas (f)	$\lvert X_i - \overline{X} \rvert$	$\lvert X_i - \overline{X} \rvert^2 \cdot f_i$
1 ⊢ 3	15	6,5	633,75
3 ⊢ 5	12	4,5	243
5 ⊢ 7	20	2,5	125
7 ⊢ 9	23	0,5	5,75
9 ⊢ 11	28	1,5	63
11 ⊢ 13	19	3,5	232,75
13 ⊢⊣ 15	19	5,5	574,75
Σ	136		1.878

$$\overline{X} = \frac{15 \cdot 2 + 12 \cdot 4 + 20 \cdot 6 + 23 \cdot 8 + 28 \cdot 10 + 19 \cdot 12 + 19 \cdot 14}{136}$$

$$\overline{X} = \frac{1156}{136}$$

$$\overline{X} = 8{,}5$$

$$S^2 = \frac{1878}{136} \;\therefore\; S^2 = 13{,}8088$$

Resposta: $S = \sqrt{S^2} \;\therefore\; S = 3{,}716$ salários mínimos

Síntese

Quando falamos que medidas de dispersão são aquelas que servem para verificar o quanto os valores encontrados em uma pesquisa estão dispersos em relação ao foco central, estamos definindo sua aplicabilidade, sem dúvida; no entanto, não podemos deixar de enfatizar o que isso representa, ou seja, é o padrão que nos permite estabelecer o grau de confiança de uma pesquisa. Nessa avaliação do grau de variabilidade do fenômeno estudado, encontramos como ferramentas de medição a amplitude total, a amplitude semi-interquartílica, o desvio médio, a variância e o desvio padrão.

Questões para revisão

1. Dado o conjunto de números

   8, 4, 6, 9, 10, 5,

   determine o desvio médio desses valores em relação à média e assinale nas opções (abaixo) o resultado.

   ( ) 12 ( ) 0 ( ) 7

   ( ) 2 ( ) 6

2. Determine a variância do conjunto de números da questão 1, supondo que esses valores correspondam a uma amostra, e marque a alternativa correta.

   ( ) 28 ( ) 7 ( ) 5,6

   ( ) 2,3664 ( ) 2,8

3. Determine o desvio padrão do conjunto de números da questão 1, supondo que esses valores correspondam a uma amostra, e assinale a alternativa correta.

   ( ) 28 ( ) 7 ( ) 5,6

   ( ) 2,3664 ( ) 2,8

4. Marque a resposta correta. Qual a amplitude total dos dados da questão 1?

   ( ) 42 ( ) 6 ( ) 4

   ( ) 3 ( ) 10

5. Determine a amplitude semi-interquartílica de uma distribuição de frequências cuja média foi 6, a mediana foi 6,5, o primeiro quartil foi 4,5 e o terceiro quartil foi 8,5. Assinale nas opções (abaixo) a alternativa correta.

   ( ) 2 ( ) 4,5 ( ) 2,5

   ( ) 6,5 ( ) 6

6. A tabela a seguir é o resultado de uma pesquisa realizada entre os funcionários de uma empresa de exportação e importação de produtos eletrônicos, com o objetivo de verificar os salários nesse segmento de mercado. Determine o desvio médio desses salários em relação à média. Calcule utilizando-se de duas casas após a vírgula.

Salários (em salários mínimos)	Funcionários
$1 \vdash 2$	1
$2 \vdash 3$	4
$3 \vdash 4$	6
$4 \vdash 5$	5
$5 \vdash 6$	6
$6 \vdash 7$	10
$7 \vdash 8$	9
$8 \vdash 9$	6
$9 \vdash 10$	3
Σ	50

( ) 1,80 ( ) 88,00 ( ) 1,76

( ) 4,33 ( ) 2,08

7. Determine a variância do conjunto de valores da questão 6, supondo que os mesmos correspondem à população total. Calcule com duas casas após a vírgula.

( ) 216,50 ( ) 4,33 ( ) 4,42

( ) 14,71 ( ) 2,08

8. Determine o desvio padrão do conjunto de valores da questão 6, supondo que os mesmos correspondem à população total. Calcule com duas casas após a vírgula.

( ) 216,50 ( ) 4,33 ( ) 4,42

( ) 14,71 ( ) 2,08

9. Qual o uso das medidas de dispersão na estatística?

10. Quais são os tipos de medidas de dispersão?

capítulo 6

Medidas de assimetria e medidas de curtose

Conteúdos do capítulo

- Medidas de assimetria
- Medidas de curtose

Após o estudo deste capítulo, você será capaz de:

1. realizar cálculos com as principais medidas de assimetria;
2. realizar cálculos com as principais medidas de curtose.

Medidas de assimetria

A média corresponde ao centro de gravidade dos dados; a variância e o desvio padrão medem a variabilidade; mas a distribuição dos pontos sobre um eixo ainda tem outras características, que podem ser medidas – uma delas é a assimetria.[1]

As MEDIDAS DE ASSIMETRIA, também denominadas de *medidas de enviesamento*, indicam o grau de deformação de uma curva de frequências.

Uma distribuição de frequências ideal seria aquela em que a curva resultante fosse rigorosamente simétrica, o que dificilmente acontece na prática. Nesse caso, a média, a mediana e a moda seriam iguais.

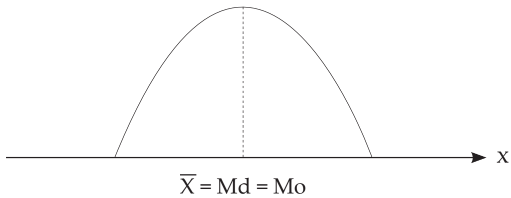
Também os quartis Q_1 e Q_3, nesse caso, ficariam equidistantes da mediana. Entretanto, se a distribuição de frequências for assimétrica, esses fatos não ocorrem.

A assimetria da curva correspondente a uma distribuição assimétrica pode apresentar uma deformidade à direita ou à esquerda. No caso de a deformidade ser à direita, a curva é dita assimétrica positiva ou desviada à direita; no caso de a deformidade ser à esquerda, a curva é dita assimétrica negativa ou desviada à esquerda.

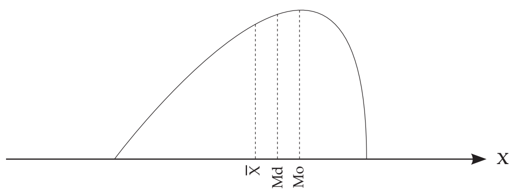
Não é comum encontrar a assimetria negativa e não é fácil identificar suas possíveis causas.[2]

A seguir, temos um exemplo gráfico de uma distribuição simétrica, uma distribuição assimétrica positiva e uma distribuição assimétrica negativa.

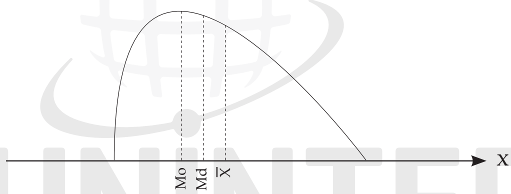
Gráfico 7 – Distribuição simétrica e assimétrica



Distribuição simétrica



Distribuição assimétrica negativa



Distribuição assimétrica positiva

Observe que, em uma distribuição assimétrica negativa, $\overline{X} < Md < Mo$ e, em uma distribuição assimétrica positiva, $\overline{X} > Md > Mo$. Na distribuição simétrica essas três medidas são iguais.

Para determinarmos o grau de assimetria de uma distribuição de frequências, são propostas várias fórmulas que nos permitem calcular o coeficiente de assimetria. Entre elas, estudaremos o coeficiente sugerido por Karl Pearson.

Pearson definiu O PRIMEIRO COEFICIENTE DE ASSIMETRIA pela fórmula:

$$As = \frac{\overline{X} - Mo}{S}$$

Como vimos no capítulo 4,

$Mo = 3 \cdot Md - 2 \cdot \overline{X}$

Então, podemos dizer que (multiplicando toda a igualdade por – 1):

$-Mo = -3 \cdot Md + 2 \cdot \overline{X}$

E somando X nos dois lados da igualdade, teremos:

$\overline{X} - Mo = -3 \cdot Md + 2 \cdot \overline{X} + \overline{X}$

Então:

$\overline{X} - Mo = -3 \cdot Md + 3 \cdot \overline{X}$

ou

$\overline{X} - Mo = 3 \cdot (\overline{X} - Md)$

Substituindo essa última expressão na fórmula de As, teremos:

$$As = \frac{3 \cdot (\overline{X} - Md)}{S}$$

Essa fórmula foi definida como o SEGUNDO COEFICIENTE DE ASSIMETRIA de Pearson.

Quando As é igual a zero, diz-se que a distribuição de frequências é simétrica; caso As seja positivo, a distribuição é assimétrica positiva ou à direita; caso As seja negativo, a distribuição é assimétrica negativa ou à esquerda.

Vamos a alguns exemplos para você entender melhor este assunto.

Exercícios Resolvidos

1. Numa distribuição de frequências que apresentou média igual a 6, mediana igual a 6 e desvio padrão igual a 1, determinar o coeficiente de assimetria dessa distribuição.

Como não conhecemos a moda, vamos utilizar a fórmula do segundo coeficiente de assimetria de Pearson.

$$As = \frac{3 \cdot (\overline{X} - Md)}{S}$$

$$As = \frac{3 \cdot (6 - 6)}{1}$$

$$As = 0$$

Resposta: Deduzimos, do resultado de As, que a distribuição é simétrica.

2. Suponhamos, agora, que uma distribuição de frequências tenha apresentado média igual a 88, mediana igual a 82 e desvio padrão igual a 40; vamos determinar o coeficiente de assimetria dessa distribuição.

Novamente utilizaremos o segundo coeficiente de assimetria de Pearson.

$$As = \frac{3 \cdot (\overline{X} - Md)}{S}$$

$$As = \frac{3 \cdot (88 - 82)}{40}$$

$$As = \frac{18}{40}$$

$$As = 0{,}45$$

Resposta: Logo, o coeficiente de assimetria dessa distribuição é 0,45.

3. Agora, vamos determinar o coeficiente de assimetria para uma distribuição que apresentou moda igual a 15, média igual a 13 e desvio padrão igual a 5.

Como não conhecemos a mediana, utilizaremos o primeiro coeficiente de assimetria de Pearson.

$$As = \frac{\overline{X} - Mo}{S}$$

$$As = \frac{13 - 15}{5}$$

$$As = -0{,}40$$

Resposta: Portanto, o coeficiente é 0,40.

Na impossibilidade de usar o desvio padrão como medida de dispersão, Pearson sugeriu outra medida de assimetria: é o coeficiente quartil de assimetria, determinado pela fórmula:

$$As = \frac{Q_1 + Q_3 - 2 \cdot Md}{Q_3 - Q_1}$$

Medidas de curtose

Diante de um conjunto de dados e desejando efetuar uma análise criteriosa deles, algumas características são importantes para podermos tomar alguma decisão. Entre elas, destacamos as medidas de posição, as medidas de dispersão e as medidas de assimetria. Entretanto, temos ainda outra ferramenta: as medidas de curtose.

A curtose é o grau de achatamento ou de afilamento de uma distribuição de frequências, ou seja, do histograma correspondente. É a medida que faltava para completarmos o quadro das estatísticas descritivas. A curtose indica o quanto uma distribuição de frequências é mais achatada ou mais afilada do que uma curva padrão, a qual é denominada de *curva normal*.

Por exemplo, se estivermos diante de dados que possuem a mesma média e a mesma amplitude total (representamos esses três conjuntos pelas Tabelas 30, 31 e 32).

Observe que, nos três casos, estamos diante de distribuições simétricas.

Tabela 30 – Conjunto de dados resultantes de uma pesquisa

Dados	Frequência
1	1
2	3
3	5
4	7
5	8
6	7
7	5
8	3
9	1

Tabela 31 – Conjunto de dados resultantes de uma pesquisa

Dados	Frequência
1	1
2	3
3	9
4	15
5	18
6	15
7	9
8	3
9	1

Tabela 32 – Conjunto de dados resultantes de uma pesquisa

Dados	Frequência
1	1
2	3
3	4
4	5
5	5
6	5
7	4
8	3
9	1

Qual a diferença entre essas três tabelas?

No primeiro caso (Tabela 30), os dados estão uniformemente distribuídos; no segundo caso (Tabela 31), os dados concentram-se em torno da média; no terceiro caso (Tabela 32), os dados têm aproximadamente a mesma frequência. Como ficariam as curvas correspondentes (histogramas) às Tabelas 30, 31 e 32, respectivamente?

Gráfico 8 – Histograma correspondente à Tabela 30

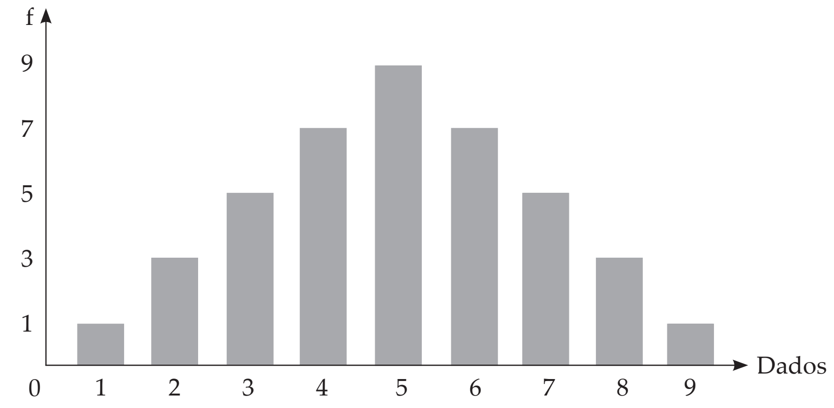


Gráfico 9 – Histograma correspondente à Tabela 31

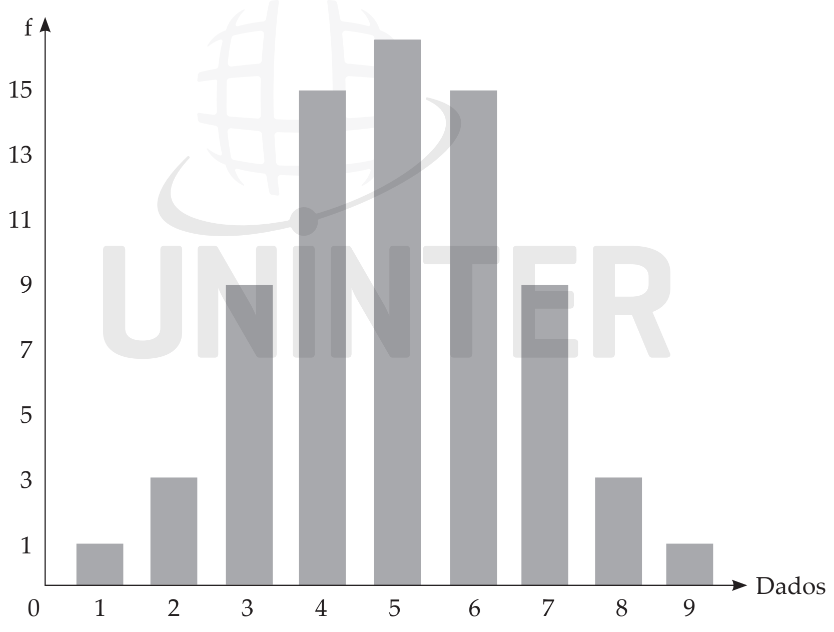
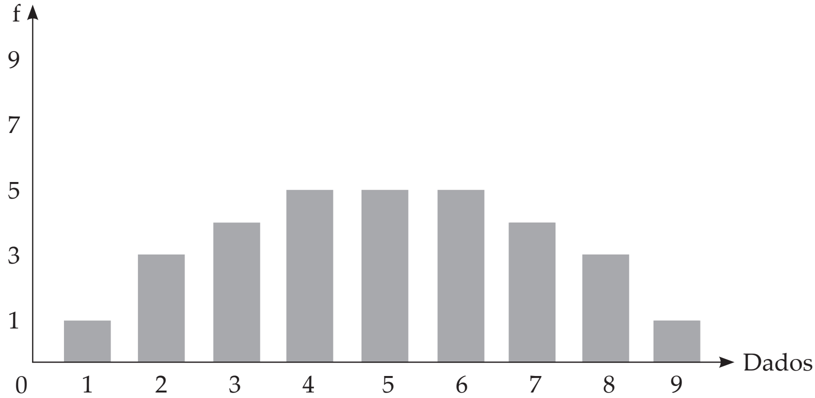


Gráfico 10 – Histograma correspondente à Tabela 32



Note que a distribuição do Gráfico 8 é normal, enquanto a distribuição do Gráfico 9 é alongada e a do Gráfico 10 é achatada.

À distribuição normal damos o nome de *curva mesocúrtica*.

À distribuição alongada damos o nome de *curva leptocúrtica*.

À distribuição achatada damos o nome de *curva platicúrtica*.

Para a determinação da curtose, sem a necessidade de fazer o gráfico, existem fórmulas que nos permitem efetuar o cálculo com maior ou com menor precisão. A mais simples, que nos dá o valor aproximado do coeficiente percentílico de curtose, é a seguinte:

$$K = \frac{Q_3 - Q_1}{2\,(p_{90} - p_{10})}$$

Em que:

K = coeficiente percentílico de curtose.

Q_1 = primeiro quartil.

Q_3 = terceiro quartil.

p_{10} = décimo percentil.

p_{90} = nonagésimo percentil.

Em que:

$$Q_1 = Li + \frac{(n/4 - \Sigma f_{ant}) \cdot A}{f_{Q_1}}$$

$$Q_3 = Li + \frac{(3.n/4 - \Sigma f_{ant}) \cdot A}{f_{Q_3}}$$

$$p_{10} = Li + \frac{(10.n/100 - \Sigma f_{ant}) \cdot A}{f_{p\,10}}$$

$$p_{90} = Li + \frac{(90.n/100 - \Sigma f_{ant}) \cdot A}{f_{p\,90}}$$

Como interpretar o resultado obtido pela utilização dessa fórmula?

É simples.

Quando K = 0,263, o achatamento da curva é igual ao da curva normal, ao que denominamos *distribuição mesocúrtica*[3]. Os dados obtidos na pesquisa, nesse caso, estão "normalmente" distribuídos.

Logo, conclui-se que, quando temos K > 0,263, estamos diante de uma curva mais achatada, ao que denominamos *distribuição platicúrtica*. Os dados obtidos na pesquisa, nesse caso, estão bem dispersos em relação à média. Trata-se, portanto, de um grupo heterogêneo.

Quando K < 0,263, estamos diante de uma curva mais alongada, que denominamos *leptocúrtica*. Os dados obtidos na pesquisa, nesse caso, estão concentrados em torno da média. Trata-se de um grupo bem homogêneo.

Vamos analisar um exemplo resolvido para ajudá-lo(a) a sanar as últimas dúvidas.

Exercício Resolvido

A Tabela 33 apresenta as faixas salariais, em número de salários mínimos, dos funcionários de determinada empresa de importação e exportação na cidade de Alegrete.

Tabela 33 – Salários dos funcionários de uma empresa

Salários	Frequência (f)
02 ⊢ 04	3
04 ⊢ 06	6
06 ⊢ 08	12
08 ⊢ 10	6
10 ⊢⊣ 12	3

Nesse caso, vamos determinar:

a) o primeiro coeficiente de assimetria de Pearson;

b) o coeficiente percentílico de curtose;

c) o tipo de curva de frequências.

Vamos, então, à solução do exemplo proposto.

a) O primeiro coeficiente de assimetria de Pearson é:

$$As = \frac{\overline{X} - Mo}{S}$$

$$\overline{X} = \frac{3 \cdot 3 + 5 \cdot 6 + 7 \cdot 12 + 9 \cdot 6 + 11 \cdot 3}{30}$$

$$\overline{X} = \frac{210}{30}$$

$$\overline{X} = 7$$

$$Mo = Li + \frac{f_{post} \cdot A}{f_{ant} + f_{post}}$$

$$Mo = 6 + \frac{6 \cdot 2}{6 + 6}$$

$$Mo = 7$$

Observe que tanto a média quanto a moda valem 7. Logo As = 0, qualquer que seja S, o que corresponde a uma distribuição de frequências simétricas.

b) O coeficiente percentílico de curtose é:

$$K = \frac{Q_3 - Q_1}{2\,(p_{90} - p_{10})}$$

Para o cálculo de Q_1, precisamos saber em que classe se encontra o elemento de ordem n/4. No caso, n/4 = 30/4 = 7,5. Logo, Q_1 encontra-se na segunda classe (intervalo de 4 a 6), pois na primeira e na segunda classes, juntas, temos nove elementos.

$$Q_1 = 4 + \frac{(7{,}5 - 3) \cdot 2}{6}$$

$$Q_1 = 4 + 1{,}5$$

$$Q_1 = 5{,}5$$

Para o cálculo de Q_3, precisamos saber em que classe encontra-se o elemento de ordem 3.n/4. No caso, 3.n/4 = 90/4 = 22,5. Logo, Q_3 encontra-se na quarta classe (intervalo de 8 a 10).

$$Q_3 = 8 + \frac{(22{,}5 - 21) \cdot 2}{6}$$

$$Q_3 = 8 + 0{,}5$$

$$Q_3 = 8{,}5$$

Para o cálculo de p_{10}, precisamos saber em que classe encontra-se o elemento de ordem 10.n/100. No caso, 10.n/100 = 10.30/100 = 3. Logo, p_{10} encontra-se na primeira classe (intervalo de 2 a 4).

$$p_{10} = 2 + \frac{(3 - 0) \cdot 2}{3}$$

$$p_{10} = 4$$

Finalmente, para o cálculo de p_{90}, precisamos saber em que classe encontra-se o elemento de ordem 90.n/100. No caso, 90.n/100 = 90.30/100 = 27. Logo, p_{90} encontra-se na quarta classe (intervalo de 8 a 10).

$$p_{90} = 8 + \frac{(27 - 21) \cdot 2}{6}$$

$p_{90} = 10$

Agora, podemos calcular o valor de K.

$$K = \frac{Q_3 - Q_1}{2\,(p_{90} - p_{10})}$$

$$K = \frac{8{,}5 - 5{,}5}{2 \cdot (10 - 4)}$$

$$K = \frac{3}{12}$$

$$K = 0{,}25$$

c) Em função do resultado obtido no item anterior, $K = 0{,}25$, verificamos que a curva é levemente leptocúrtica ($K < 0{,}263$).

Síntese

Para realizar uma análise criteriosa de dados, você deve recorrer a várias ferramentas, entre elas as estudadas neste capítulo. Como vimos, as medidas de assimetria, ou de enviesamento, são fundamentais na estatística, pois indicam o grau de deformação de uma curva de frequências. Na distribuição de frequências ocorre, de acordo com a direção (esquerda ou direita) da deformidade, ou desvio da curva, a denominada *assimétrica negativa* ou *assimétrica positiva*. Temos, portanto, a distribuição simétrica (quando a média, a mediana e a moda são iguais, sendo que os quartis Q_1 e Q_3 ficam equidistantes da mediana), a assimétrica negativa (com desvio à esquerda) e a assimétrica positiva (com desvio à direita). Mas como medimos o grau de assimetria? Como você pôde observar nos exercícios resolvidos, utilizamos nessa tarefa fórmulas como: a do segundo coeficiente de Karl Pearson e também a do coeficiente quartil de assimetria.

No entanto, nesse processo de medidas não ficamos apenas com as de assimetria. Utilizamos também as medidas de posição, as medidas de dispersão e as de curtose. Esta última, a de curtose, é útil para medir o grau de achatamento ou de afilamento de uma distribuição de frequências.

Questões para revisão

1. Em uma distribuição de frequências, verificamos que a moda é igual a 8,0, a média é igual a 7,8 e o desvio padrão é igual a 1,0. Determine o coeficiente de assimetria de Pearson e assinale a resposta correta na relação abaixo.

   ( ) 0,20 ( ) 2,0 ( ) 0,50

   ( ) – 0,20 ( ) – 2,0

2. Em uma distribuição de frequências, verificou-se que a mediana é igual a 15,4, a média é igual a 16,0 e o desvio padrão é igual a 6,0. Determine o coeficiente de assimetria de Pearson e assinale a resposta correta na relação abaixo.

   ( ) 0,10 ( ) 0,30 ( ) 0,50

   ( ) – 0,10 ( ) – 0,30

3. Observamos que, em uma distribuição de frequências, o primeiro quartil é igual a 3, o terceiro quartil é igual 8, o décimo centil é igual a 1,5 e o nonagésimo centil é igual a 9. Com base nesses resultados, podemos afirmar que se trata de uma curva:

   ( ) mesocúrtica, com k = 0,263;

   ( ) leptocúrtica, com k = 0,233;

   ( ) leptocúrtica, com k = 0,25;

   ( ) platicúrtica, com k = 0,45;

   ( ) platicúrtica, com k = 0,333.

4. O coeficiente de curtose (K) para uma determinada distribuição de frequências é igual a 0,297. Podemos, então, afirmar que a curva é:

   ( ) mesocúrtica;

   ( ) platicúrtica;

   ( ) leptocúrtica;

   ( ) assimétrica positiva;

   ( ) simétrica.

5. O segundo coeficiente de assimetria de Pearson, para determinada distribuição de frequências, é igual a zero. Podemos, então, afirmar que a curva é:

   ( ) mesocúrtica;

   ( ) leptocúrtica;

   ( ) platicúrtica;

   ( ) simétrica;

   ( ) assimétrica positiva.

6. Descreva o que indicam as medidas de assimetria.

7. O que indicam as medidas de curtose?

capítulo 7

Cálculo de probabilidades

Conteúdos do capítulo

- Definições e conceitos básicos da teoria das probabilidades
- Cálculos de probabilidade
- Aplicação da regra de adição, da de multiplicação e do teorema de Bayes em estatística

Após o estudo deste capítulo, você será capaz de:

1. conceituar a teoria das probabilidades;
2. aplicar em casos práticos os conceitos teóricos;
3. aplicar a regra da adição, a da multiplicação e o teorema de Bayes em casos de cálculos da probabilidade.

Teoria das probabilidades: conceitos básicos

Probabilidade, num conceito amplo, é o estudo dos fenômenos aleatórios.

Os fenômenos estudados em estatística são fenômenos cujo resultado, mesmo em condições normais de experimentação, variam de uma observação para outra, dificultando dessa maneira a previsão de um resultado futuro. Para a explicação desses fenômenos, chamados de *aleatórios*, adota-se o modelo matemático denominado *teoria das probabilidades*.

Historicamente, a teoria da probabilidade começou com o estudo dos jogos de azar, como a roleta e as cartas[1]. Hoje, suas aplicações são inúmeras.

Para entender a inferência estatística, é preciso entender vários conceitos de probabilidade. A probabilidade e a estatística estão estreitamente relacionadas, porque formulam tipos opostos de questões. Na probabilidade, sabemos como um processo ou experimento funciona e queremos predizer quais serão os resultados de tal processo. Em estatística, não sabemos como um processo funciona, mas podemos observar os seus resultados e utilizar as informações sobre os mesmos para conhecer a natureza do processo ou do experimento.

O termo *probabilidade* é usado de modo muito amplo na conversação diária para sugerir um certo grau de incerteza sobre o que ocorreu no passado, o que ocorrerá no futuro e o que está ocorrendo no presente.

O torcedor de certo time pode apostar contra ele porque sua probabilidade de ganhar é pequena. O serviço de meteorologia pode prever tempo ruim porque a probabilidade de chuvas é grande.

É também importante o cálculo das probabilidades em situações que envolvam tomada de decisão. Um empresário, por exemplo, precisa de informações sobre a probabilidade de sucesso para a exportação de determinado produto em um determinado mercado.

Logo, os modelos probabilísticos podem ser úteis em diversas áreas do conhecimento humano, tais como em administração de empresas, ciências econômicas ou comércio exterior.

Experimento aleatório E é aquele que poderá ser repetido sob as mesmas condições indefinidamente. Tal experimento apresenta variações de resultados, não sendo possível afirmar *a priori* qual será sua determinação antes que o mesmo tenha sido realizado. É possível, porém, descrever a probabilidade de cada possível resultado, bem como é possível relacionar todos os resultados que podem ocorrer. O lançamento de uma moeda é um experimento aleatório, pois esse experimento poderá ser repetido quantas vezes desejarmos. Antes do lançamento, não poderemos dizer qual será o resultado, mas somos capazes de relatar os possíveis resultados: cara ou coroa.

Assim, a estatística baseia-se em experimentos enquanto o cálculo de probabilidades baseia-se em postulados lógicos. O cálculo das probabilidades e a estatística estão relacionados por intermédio da chamada *Lei dos Grandes Números*[2], que será estudada adiante.

Espaço amostral

Definimos espaço amostral S como sendo o conjunto de todos os possíveis resultados de um experimento E.

Por exemplo, no experimento que consiste no lançamento de um dado, o espaço amostral é S = {1, 2, 3, 4, 5, 6}.

Evento

Evento é qualquer conjunto de resultados de um experimento. Sendo o evento um subconjunto de S, indicaremos os eventos por letras maiúsculas: A, B, C, e assim por diante.

Por exemplo, no experimento que consiste no lançamento de um dado, podemos definir o evento A como "sair um número par".

Então, A = {sair número par}.

A = {2, 4, 6}.

Evento Simples

É aquele formado por um único elemento do espaço amostral.

Por exemplo, no experimento que consiste no lançamento de um dado, o evento A, definido como "sair um número maior que 5", é um evento simples.

A = {6}.

Evento composto

É aquele que possui mais de um elemento. Note que o evento A = {2, 4, 6} é composto.

Evento certo e evento impossível

Diante das explicações sobre eventos, notamos que S (espaço amostral) e ∅ (conjunto vazio) também são eventos, chamados respectivamente de *evento certo* e *evento impossível.*

Definição de probabilidade

A fim de entendermos probabilidade, vamos utilizar processos aleatórios familiares, tais como:

a) jogar uma moeda vinte vezes e observar o número de caras obtidas;

b) retirar uma carta de um baralho comum de 52 cartas e observar seu naipe;

c) jogar um dado dez vezes e observar quantas vezes a face que contém o 6 é voltada para cima.

O conhecimento de como se aplicam os princípios da probabilidade às situações em que sabemos o modo como o processo funciona vai nos ajudar a compreender de que maneira podemos usar a inferência estatística para conhecer a natureza de um processo desconhecido.

A análise dos três processos anteriormente exemplificados revela-nos que:

a) cada experimento poderá ser repetido indefinidamente sob as mesmas condições;

b) não conhecemos, a princípio, um resultado particular do experimento, mas sabemos todos os possíveis resultados;

c) quando o experimento é repetido um grande número de vezes, surge uma regularidade, ou seja, podemos observar uma tendência. Essa regularidade é observada pela estabilidade da fração fr = f/n (frequência relativa), em que f é o número de sucessos de um resultado particular estabelecido antes do experimento e n é o número de repetições.

Suponhamos, agora, duas outras situações. Na primeira, extraímos aleatoriamente 10 (dez) bolas de uma caixa que contém 20 (vinte) bolas pretas e 80 (oitenta) bolas vermelhas. Pela teoria das probabilidades, podemos determinar que existe a probabilidade de exatamente oito dessas bolas serem vermelhas e as outras duas serem pretas. A probabilidade é dada por um valor numérico entre 0 e 1. No exemplo dado, obter oito bolas vermelhas e duas pretas é o que chamamos de *esperança matemática*, ou seja, o resultado com maior probabilidade de ocorrer. Na segunda, antes de uma eleição presidencial, perguntamos a 2.000 pessoas escolhidas aleatoriamente o nome do seu candidato e, utilizando a inferência estatística, é possível prever as tendências presidenciais da população, tendo como base os resultados obtidos na amostra.

A probabilidade matemática de um acontecimento é a relação entre o número de casos favoráveis e o número de casos possíveis, desde que haja rigorosa equipossibilidade entre todos os casos.

Designando por *S* o número de casos possíveis e por *A* o número de casos favoráveis, temos a probabilidade P, assim definida:

$$P(A) = \frac{A}{S}$$

Ou seja:

$$P(A) = \frac{\text{número de elementos do evento A}}{\text{número de elementos do espaço amostral S}}$$

O valor de P(A) é sempre uma fração compreendida entre zero e um, pois o número de casos favoráveis nunca pode ser maior que o número de casos possíveis.

Os valores limites da probabilidade são:

a) P(A) = 0 quando A = 0, isto é, não há casos favoráveis; há certeza de não acontecer;

b) P(A) = 1 quando A = S, isto é, todos os casos são favoráveis, havendo certeza do acontecimento.

A probabilidade do não acontecimento costuma ser simbolizada pela letra Q, em que Q(A) = 1 – P(A).

Logo: P(A) + Q(A) = 1

Lei dos grandes números

Conforme DuPasquier, citado por Farias, Soares e Cesar[3], em uma série de observações de um conjunto natural, realizadas todas em circunstâncias idênticas, um atributo "x" ocorre com uma frequência relativa, cujo valor é uma aproximação da probabilidade, aproximação essa tanto maior quanto maior for o número de observações.

Por exemplo, qual a probabilidade de obtermos uma cara em uma única jogada de uma moeda honesta?

S = {cara, coroa}

A = {deu cara}

P(A) = 1/2 ou 50%

Ou, então, qual a probabilidade de obtermos uma única cara em um lançamento de três moedas?

Considerando K = cara , C = coroa e sendo A = {deu uma única cara}, temos:

S = {KCC, KKC, KKK, CKC, CKK, CCK, CCC, KCK}

Então, P(A) = 3/8.

Ainda qual a probabilidade de obtermos como resultado o 5 em uma única jogada de um dado honesto?

A = {deu 5}

S = {1 , 2 , 3 , 4 , 5 , 6}

P(A) = 1/6

Curiosidade: na estatística, uma ferramenta honesta é aquela que não é viciada. Por exemplo, um dado honesto significa um dado que não é viciado. Em outras palavras, é um dado em que qualquer resultado tem a mesma probabilidade de ocorrer, para qualquer face.

Exercícios Resolvidos

1. Qual a probabilidade de obtermos o total de seis (6) pontos na jogada de dois (2) dados honestos?

S = {36 resultados possíveis}

A = {a soma dos dois dados é igual a 6}

A = {(1 , 5), (2 , 4), (3 , 3), (4 , 2), (5 , 1)}

P(A) = 5/36

2. Qual a probabilidade de sair uma figura (valete, dama ou rei) na retirada de uma única carta de um baralho comum de 52 cartas?

S = {52 resultados possíveis}

A = {a carta retirada é uma figura}

P(A) = 12/52 = 3/13

Obs.: lembrar que temos quatro valetes, quatro damas e quatro reis.

3. Qual a probabilidade de acertarmos o resultado do sorteio da mega-sena com um único cartão jogado de seis dezenas?

O número de resultados possíveis, S, é igual a $C_{60,6}$.

Como $C_{N,X} = \dfrac{N!}{X!\,(N-X)!}$, temos que

$$C_{60,6} = \frac{60!}{6!\,(60-6)!}$$

$C_{60,6}$ = 50.063.860

Seja A = {acertar as 6 dezenas sorteadas}

$$P(A) = \frac{1}{50.063.860} = 0,000.000.019 \text{ ou } 0,000.001.9\%.$$

Esperança matemática

Aos valores que em estatística denominamos de *médias*, no cálculo das probabilidades denominamos de *esperanças matemáticas*.

O cálculo da esperança é possível pela fórmula:

$E(x) = n \cdot p$

Em que:

> $E(x)$ = esperança matemática de ocorrer o evento x;
> n = número de tentativas;
> p = probabilidade de sucesso, ou seja, probabilidade de ocorrer o evento x em uma tentativa única.

Originalmente, o conceito de esperança matemática surgiu em função de sua relação com jogos de azar e, em sua forma mais simples, é o produto da quantia que um jogador pode ganhar (o prêmio) pela probabilidade de que isso ocorra.

Assim, se em uma caixa encontram-se 100 canetas, sendo 20 pretas, 50 vermelhas e 30 azuis, qual o valor esperado de canetas azuis, em 20 retiradas de uma caneta, se houve a reposição da mesma na caixa?

$$p = \frac{30}{100}$$

$$E(x) = n \cdot p$$

$$E(x) = 20 \cdot \frac{30}{100}$$

$$E(x) = 6$$

Ou seja, esperamos que, nas vinte tentativas, seis canetas sejam azuis.

Exercício Resolvido

Qual a esperança matemática de uma pessoa que joga na aposta mínima da mega-sena, cujo prêmio previsto para o ganhador é de R$ 1.000.000,00?

Já vimos, no exemplo 6, que o número de resultados possíveis é de 50.063.860, e a probabilidade de ganhar o prêmio é de $\frac{1}{50.063.860}$

A esperança matemática é:

$E(x) = n \cdot p$

$E(x) = 1.000.000 \cdot \frac{1}{50.063.860}$

$E(x) = 0,02$

O que esse resultado nos mostra? Mostra que só devemos pagar R$ 0,02 por aposta, um valor muito inferior ao que é cobrado da pessoa que pretende "tentar a sorte".

Probabilidades finitas dos espaços amostrais finitos

Seja S um espaço amostral finito, em que $S = \{a_1, a_2, ..., a_n\}$. Considere o evento formado por um resultado simples: $A = \{a_i\}$.

A cada evento simples $\{a_i\}$, associa-se um número P_i, denominado de *probabilidade de $\{a_i\}$ ocorrer*, satisfazendo as seguintes condições:

a) $P_i \geq 0$, sendo i = 1, 2, ..., n;

b) $P_1 + P_2 + ... + P_n = 1$.

A probabilidade P(A) de cada evento composto (mais de um elemento) é então definida pela soma das probabilidades dos pontos de A.

Portanto, se três cavalos – A, B e C – estão em uma corrida, e o cavalo A tem duas vezes mais probabilidade de ganhar que o B e o cavalo B tem duas vezes mais probabilidade de ganhar que o C, quais são as probabilidades de vitória de cada cavalo, isto é, P(A), P(B) e P(C)?

Faça P(C) = p; então, P(B) = 2.p e P(A) = 4.p. Como a soma das probabilidades é igual a 1, temos:

$p + 2.p + 4.p = 1$ ou $7.p = 1$ ou $p = 1/7$.

Então:

P(A) = 4/7

P(B) = 2/7

P(C) = 1/7

Espaços amostrais finitos equiprováveis

Quando associamos a cada ponto amostral a mesma probabilidade de ocorrência, o espaço amostral é dito *equiprovável* ou *uniforme*.

Em particular, se S contém n pontos, então a probabilidade de cada ponto é 1/n.

Logo, se você escolher aleatoriamente (a expressão *aleatória* nos indica que o espaço é equiprovável) uma carta de um baralho comum com 52 cartas, em que seja A = {a carta é de ouros} e B = {a carta é uma figura}, deve calcular P(A) e P(B).

$$P(A) = \frac{\text{número de ouros}}{\text{total de cartas}} = \frac{13}{52} = \frac{1}{4}$$

$$P(B) = \frac{\text{número de figuras}}{\text{total de cartas}} = \frac{12}{52} = \frac{3}{13}$$

Acontecimentos mutuamente exclusivos

Dois acontecimentos podem ser mutuamente exclusivos, isto é, ocorrendo um deles não pode ocorrer o outro. Por exemplo, se ao lançarmos um dado aparecer o número 4, então não pode ter aparecido o número 5.

Regra da adição para eventos mutuamente exclusivos

Quando A e B são eventos mutuamente exclusivos, $P(A \cup B) = P(A) + P(B)$.

Como nós lemos essa igualdade?

Lemos: a probabilidade de acontecer o evento A ou o evento B é igual à probabilidade de acontecer o evento A mais a probabilidade de acontecer o evento B.

O símbolo ∪ corresponde, na teoria dos conjuntos, à união e, na aritmética, à soma. Corresponde, ainda, na álgebra boolena, à disjunção "OU".

Vamos verificar, no exemplo do cálculo de "esperança matemática", na página 106, qual seria a probabilidade de B ou C ganhar?

$P(B \cup C) = P(B) + P(C) = 2/7 + 1/7 = 3/7.$

Ou, então, qual a probabilidade de tirarmos o primeiro prêmio em um sorteio de 100 bilhetes, estando-se de posse de três desses bilhetes?

A probabilidade de cada bilhete é de 1/100. Logo, aplicando a regra da adição, a probabilidade de ganharmos é: $P = 1/100 + 1/100 + 1/100 = 3/100$.

Eventos não mutuamente exclusivos

Consideremos o caso de dois eventos que podem ocorrer simultaneamente. Suponhamos A = {extração de um ás de um baralho} e B = {extração de uma carta de espadas}.

Então, A e B não são mutuamente exclusivos, visto que pode ser extraído o ás de espadas.

Somando simplesmente $P(A) + P(B)$, estamos contando duas vezes os resultados de $(A \cap B)$. Por conseguinte, para obter o número correto de resultados $P(A \cup B)$, devemos subtrair $P(A \cap B)$.

O símbolo ∩ corresponde, na teoria dos conjuntos, à interseção e, na aritmética, à multiplicação. Corresponde, ainda, na álgebra booleana, à conjunção "E".

Regra da adição para eventos não mutuamente exclusivos

Quando A e B são eventos não mutuamente exclusivos,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, em que

$P(A \cap B) = P(A) \cdot P(B)$

A fórmula $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ é a regra geral de adição, sejam os eventos mutuamente exclusivos ou não. Por quê? Porque, no caso dos eventos serem mutuamente exclusivos, $P(A \cap B) = 0$.

Devemos estar atentos para o fato de que $P(A \cap B) = P(A) \cdot P(B)$, quando os eventos A e B são independentes.

Quando um evento A é independente de um evento B, o evento B também é independente do evento A.

Por exemplo, ao jogarmos uma moeda, se definirmos o evento A como "deu cara" e o evento B como "deu coroa", é fácil observar que se "deu cara" não pode ter "dado coroa" e vice-versa.

Logo, dois eventos são independentes quando a ocorrência de um deles não depende da ocorrência do outro.

Há eventos, entretanto, que são dependentes.

Por exemplo, em um carregamento de mil sacas de café, 140 delas estão com peso abaixo do recomendado. Se uma pessoa escolher ao acaso duas sacas desse carregamento, qual é a probabilidade das duas estarem abaixo do peso?

Nesse caso, ao retirarmos a segunda saca, devemos levar em conta que a primeira já havia sido retirada. Logo, para o cálculo da probabilidade procurada, esse fato deve ser levado em consideração. Você aprenderá a fazer esse cálculo logo adiante, ao estudar probabilidade condicional.

Exercícios Resolvidos

1. Ao se retirar uma carta de um baralho comum de 52 cartas, qual a probabilidade de ela ser ou um ás ou uma carta de espadas?

No presente caso, $P(A) = 4/52$, $P(B) = 13/52$ e $P(A \cap B) = 4/52 \times 13/52 = 1/52$.

Então, $P(A \cup B) = 4/52 + 13/52 - 1/52 = 16 / 52 = 4/13$.

2. Jogamos um dado branco e um dado preto. Calcule a probabilidade de ocorrer:

a) soma igual a 5;

b) soma igual a 11.

Sabemos que, ao jogar dois dados, existem 36 diferentes resultados (os seis (6) do primeiro dado, vezes os seis (6) do segundo dado). Então:

S = {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

a) A soma igual a 5 pode ocorrer nos seguintes casos:

A = {(1, 4), (2, 3), (3, 2), (4, 1)}

Sabemos, pela definição de probabilidade, que:

$$P(A) = \frac{\text{número de elementos do evento A}}{\text{número de elementos do espaço amostral S}}$$

Então, temos:

$$P(A) = \frac{4}{36} = 0{,}1111 \text{ ou } 11{,}11\%$$

b) A soma igual a 11 pode ocorrer nos seguintes casos:

B = {(5, 6), (6, 5)}

Agora, temos:

$$P(B) = \frac{2}{36} = 0{,}0556 \text{ ou } 5{,}56\%$$

3. Quando uma carta é retirada de um baralho comum de 52 cartas, QUAL A PROBABILIDADE (CALCULE) DE:

a) sair uma carta vermelha;

b) sair uma carta de copas;

c) sair um valete.

Sabendo que um baralho comum tem 52 cartas, assim distribuídas:

- naipe de copas: 13 cartas – estas são cartas vermelhas;
- naipe de ouros: 13 cartas – estas são cartas vermelhas;
- naipe de paus: 13 cartas – estas são cartas pretas;
- naipe de espadas: 13 cartas – estas são cartas pretas.

a) Ao retirarmos desse baralho uma única carta, a probabilidade de ela ser uma carta vermelha é:

A = {carta é vermelha}

$$P(A) = \frac{26}{52} = 0{,}5 \text{ ou } 50\%$$

b) Ao retirarmos desse baralho uma única carta, a probabilidade de ela ser uma carta de copas é:

B = {carta é de copas}

$$P(B) = \frac{13}{52} = 0{,}25 \text{ ou } 25\%$$

c) Ao retirarmos desse baralho uma única carta, a probabilidade de ela ser um valete é:

C = {carta é um valete}

$$P(C) = \frac{4}{52} = 0{,}0769 \text{ ou } 7{,}69\%$$

4. Em uma disputa final de torneio de tiro ao alvo, a probabilidade de Kendric acertar no alvo é de 1/2 e a de Marcel atingir o mesmo alvo é de 3/5. Qual a probabilidade do alvo ser atingido, se ambos atirarem nele?

Estamos diante de dois eventos NÃO mutuamente exclusivos, ou seja, se Kendric acertar o alvo, não exclui a possibilidade de Marcel também o acertar. Logo, se definirmos que:

A = {Kendric acertou o alvo} e

B = {Marcel acertou o alvo},

temos que:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Em que:

$$P(A \cap B) = P(A) \cdot P(B)$$

Ou seja:

$$P(A \cup B) = \frac{1}{2} + \frac{3}{5} - \frac{1}{2} \cdot \frac{3}{5}$$

$$P(A \cup B) = \frac{1}{2} + \frac{3}{5} - \frac{3}{10}$$

$$P(A \cup B) = \frac{8}{10} = 0{,}80 \text{ ou } 80\%$$

Acontecimentos simultâneos e sucessivos

Há acontecimentos independentes que podem ser simultâneos, como o caso de um baralho em que queremos tirar uma carta de espadas e uma carta que seja uma figura. A isso damos o nome de ACONTECIMENTO COMPOSTO.

Há, no entanto, acontecimentos independentes sucessivos. São os que exigem a ocorrência do primeiro fato para que possa existir a hipótese do segundo acontecimento, tal como neste exemplo: em duas jogadas sucessivas de um dado, qual a probabilidade que temos de sair o número 5 as duas vezes? A isso damos o nome de ACONTECIMENTO COMPLEXO.

Probabilidade condicional

Dados dois eventos, A e B, de um espaço amostral S, denota-se por $P(A \mid B)$ a probabilidade condicionada de ocorrer o evento A quando o evento B já tiver ocorrido, assim:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ com } P(B) \neq 0, \text{ pois B já ocorreu.}$$

Então, para avaliar a probabilidade de ocorrer A, dado que B já ocorreu, basta contar o número de casos favoráveis do evento $(A \cap B)$ e dividir esse número pela quantidade de casos favoráveis do evento B.

Quando os eventos A e B são independentes, temos $P(A \mid B) = P(A)$ ou $P(B \mid A) = P(B)$.

Vamos analisar alguns exemplos para que você possa entender melhor o que foi dito.

Caso seja E o experimento que consiste em lançar um dado, e o evento A = {saiu o número 2}, então, $P(A) = 1/6$.

Considere agora o evento B = {saiu um número par}.

Então, $P(B) = 3/6 = 1/2$.

Nesse exemplo, podemos estar interessados em avaliar a probabilidade de que ocorra o evento A (saiu o número 2), condicionada à ocorrência do evento B (saiu um número par). Em símbolos, designa-se por $P(A \mid B)$.

Observe que, dada a informação da ocorrência do evento B, reduzimos o espaço amostral S de {1, 2, 3, 4, 5, 6} para {2, 4, 6}. Logo $P(A \mid B) = 1/3$.

Ou, se for o caso de um número ser sorteado ao acaso entre os inteiros de 1 a 15. Se o número sorteado for par, QUAL A PROBABILIDADE DE QUE SEJA O NÚMERO 6?

S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}

Chamemos A = {o número é 6},

e B = {o número é par}.

Então, $P(A \mid B) = ?$

Note que a probabilidade de ocorrência do evento A, sem a informação da ocorrência do evento B, é:

$P(A) = 1/15$

Dada, porém, a informação de que o número sorteado foi par, o espaço amostral reduz-se para {2, 4, 6, 8, 10, 12, 14} e é nesse espaço amostral que iremos avaliar a probabilidade de ocorrer o evento A.

$$P(A \mid B) = \frac{\text{número de casos favoráveis do evento } A \cap B}{\text{número de casos favoráveis do evento } B} = \frac{1}{7}$$

Pois $(A \cap B) = \{ 6 \}$ e B = {2, 4, 6, 8, 10, 12, 14}.

E, em uma situação em que se retira, sem reposição, duas peças de um lote de dez peças, em que apenas quatro são boas, QUAL A PROBABILIDADE DE QUE AMBAS SEJAM DEFEITUOSAS?

A = {a primeira peça é defeituosa}

B = {a segunda peça é defeituosa}

$P(A \cap B) = P(A) \cdot P(B \mid A)$

$P(A \cap B) = 6/10 \cdot 5/9 = 1/3$

Observe que $P(A \mid B)$ é a probabilidade da segunda peça ser defeituosa, dado que a primeira era defeituosa.

> Observe que, em experimentos em que ocorre reposição, o elemento escolhido é devolvido à população, podendo ser escolhido novamente. Se não houver reposição, o elemento, uma vez escolhido, não é devolvido à população, não podendo, assim, ser escolhido novamente.

Regra da multiplicação

A partir do que vimos na probabilidade condicional, tanto no acontecimento composto como no acontecimento complexo, aplica-se a regra da multiplicação, que diz:

> A probabilidade de um acontecimento complexo ou composto é igual ao produto das probabilidades simples dos acontecimentos considerados (quando os acontecimentos anteriores já se realizaram).

Ou seja:

$P(A \cap B) = P(B) \cdot P(A \mid B)$ ou $P(A \cap B) = P(A) \cdot P(B \mid A)$

Quando os eventos A e B são independentes, temos que a $P(A \cap B) = P(A) \cdot P(B)$

O princípio da multiplicação afirma que, se o primeiro de dois experimentos admite "a" resultados possíveis e o segundo comporta "b" resultados possíveis, podendo ocorrer qualquer combinação desses resultados, então o número total de resultados possíveis dos dois experimentos é $a \cdot b$

Vamos ver como isso funciona.

Suponhamos que temos cinco candidatos de direita para disputar as eleições à presidência, e seis candidatos de esquerda disputando as mesmas eleições. Qual o número de pares possíveis de candidatos na eleição geral, sendo um de direita e outro de esquerda?

Número de candidatos = $6 \cdot 5 = 30$

Agora, em três jogadas de um dado, qual a probabilidade de sair três vezes seguidas o número 2?

Na primeira jogada, a probabilidade é de 1/6.

Na segunda jogada, a probabilidade é de 1/6.

Na terceira jogada, a probabilidade é de 1/6.

Então, a probabilidade procurada é, pela regra da multiplicação, igual a:

$P = 1/6 \cdot 1/6 \cdot 1/6 = 1/216$

Vamos transformar esse resultado em porcentagem:

$\frac{1}{216} = 0{,}00463$, ou seja, 0,463%

Exercícios Resolvidos

1. Em uma caixa, temos três bolas brancas, duas bolas pretas e cinco bolas amarelas. Qual a probabilidade de retirarmos duas bolas brancas, uma após a outra, sem reposição?

A = {a primeira bola é branca}

B = {a segunda bola é branca}

$P(A \cap B) = P(A) \cdot P(B \mid A)$

$P(A \cap B) = 3/10 \cdot 2/9 = 1/15$

Vamos transformar esse resultado em porcentagem:

$\frac{1}{15} = 0{,}0667$, ou seja, 6,67%

2. Num lote de doze peças, quatro são defeituosas. Nesse caso, considerando que três peças são retiradas aleatoriamente, uma após a outra, sem reposição, encontre a probabilidade de todas essas peças retiradas serem NÃO defeituosas.

A probabilidade de a primeira peça ser não defeituosa é 8/12, já que oito das doze peças são não defeituosas. Se a primeira peça é não defeituosa, então a probabilidade de a próxima ser não defeituosa é 7/11, pois somente sete das onze peças restantes são não defeituosas. Finalmente, se a segunda peça também é não defeituosa, a probabilidade de a terceira ser não defeituosa é 6/10, já que somente seis das dez peças restantes são não defeituosas. Então:

A = {a primeira peça é não defeituosa}

B = {a segunda peça é não defeituosa}

C = {a terceira peça é não defeituosa}

$P(A \cap B \cap C) = 8/12 \cdot 7/11 \cdot 6/10 = 14/55$

Independência estatística

Um evento A é considerado independente de um evento B se a probabilidade de A é igual à probabilidade condicional de A, dado B, isto é, se:

$P(A) = P(A \mid B)$.

É evidente que se A é independente de B, e B é independente de A, temos:

$P(B) = P(B \mid A)$.

Considerando-se a regra da multiplicação, podemos afirmar que se A e B são independentes, então:

$P(A \cap B) = P(A) \cdot P(B)$.

Dados n eventos $A_1, A_2, \ldots, A_n$, dizemos que eles são independentes se o forem 2 a 2, 3 a 3, ... , n a n.

Então: $P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2) \ldots P(A_n)$.

Vamos analisar alguns exemplos.

No primeiro, vamos considerar duas pessoas identificadas como A e B. Nesse caso, se a probabilidade de que A esteja viva daqui a 20 anos é 70% e a de que B esteja viva daqui a 20 anos é de 50%, então A PROBABILIDADE DE QUE AMBAS ESTEJAM VIVAS DAQUI A 20 ANOS É DE:

P(A e B estarem vivas) = 70/100 . 50/100 = 35/100 ou 35%

Como segundo caso, analisaremos o lançamento de dois dados. Verifique se os eventos A e B são independentes, sendo que A é o evento em que a soma dos pontos dos dois dados é quatro (4), e B o evento em que no primeiro dado deu 1. Lembre-se que, para essa experiência, o espaço amostral tem 6 x 6 = 36 elementos, dos quais seis correspondem a pares de números em que no primeiro dado deu 1.

S = {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

Logo, P(B) = 6/36 = 1/6

Os elementos do espaço amostral que indicam soma igual a quatro (4) são:

(1, 3) , (2, 2) e (3, 1).

Logo, P(A) = 3/36 = 1/12.

Então, a probabilidade de a soma ter sido quatro (4), se sabemos que no primeiro dado deu o número 1, é:

$P(A \mid B) = 1/3$

Evidentemente, A e B não são independentes, pois P(A) é diferente de $P(A \mid B)$.

Exercícios Resolvidos

1. Uma urna contém oito bolas brancas, sete bolas pretas e quatro bolas verdes. Uma bola é retirada aleatoriamente dessa urna. CALCULE A PROBABILIDADE DE:

a) sair uma bola branca;

b) sair uma bola preta;

c) sair uma bola verde.

Vamos calcular inicialmente a probabilidade de a bola ter sido branca. Como temos oito bolas brancas de um total de dezenove bolas (8 + 7 + 4), a probabilidade procurada é:

$$P\text{ (bola ser branca)} = \frac{8}{19}$$

Agora, a probabilidade de a bola ter sido preta. Como temos sete bolas pretas de um total de dezenove bolas, a probabilidade procurada é:

$$P\text{ (bola ser preta)} = \frac{7}{19}$$

Finalmente, vamos calcular a probabilidade de a bola ter sido verde. Temos quatro bolas verdes. Então:

$$P\text{ (bola ser verde)} = \frac{4}{19}$$

2. Um número inteiro é escolhido ao acaso dentre os números de 1 a 30. Calcule a probabilidade de:

a) o número ser divisível por 3;

b) o número ser divisível por 5;

c) o número ser divisível por 3 ou por 5;

d) o número ser divisível por 3 e por 5.

> É necessário agora relembrar que associamos a operação lógica OU à operação aritmética + (soma) e associamos a operação lógica E à operação aritmética x (multiplicação).

Vamos, então, à solução do exemplo 27.

S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30}

Chamemos de A = {o número é divisível por 3}

Então:

$P(A) = \frac{10}{30} = \frac{1}{3}$, pois temos dez números divisíveis por 3.

Chamemos de B = {o número é divisível por 5}

$P(A) = \frac{6}{30} = \frac{1}{5}$, pois temos seis números divisíveis por 5.

Chamemos de C = {o número é divisível por 3 ou por 5}

Então:

$P(C) = P(A) + P(B) - P(A) \cdot P(B)$

$$P(C) = \frac{1}{3} + \frac{1}{5} - \frac{1}{3} \cdot \frac{1}{5}$$

$$P(C) = \frac{1}{3} + \frac{1}{5} - \frac{1}{15}$$

$$P(C) = \frac{5 + 3 - 1}{15}$$

$$P(C) = \frac{7}{15}$$

Chamemos de D = {o número é divisível por 3 e por 5}

Então:

$$P(D) = \frac{1}{3} \cdot \frac{1}{5}$$

$$P(D) = \frac{1}{15}$$

3. Uma urna contém cinco bolas brancas e três vermelhas, sendo que uma outra urna contém quatro bolas brancas e cinco vermelhas. Considerando que uma bola é retirada de cada urna, ENCONTRE A PROBABILIDADE DE SEREM:

a) da mesma cor;

b) de cores diferentes.

Para facilitar a visualização do exercício, vamos representar as duas urnas:

Urna 1	**Urna 2**
5 brancas	4 brancas
3 vermelhas	5 vermelhas

Calculemos então a probabilidade de as duas bolas, uma retirada da urna 1 e outra retirada da urna 2, serem ambas da mesma cor.

As duas poderão ser brancas. Então:

$$P\text{ (duas brancas)} = \frac{5}{8} \cdot \frac{4}{9}$$

Observe que multiplicamos as probabilidades, porque estamos supondo que uma E outras bolas são brancas. Portanto, à operação lógica E, associamos a multiplicação, lembra-se?

Mas as duas poderiam ter sido vermelhas. Então:

$$P\text{ (duas vermelhas)} = \frac{3}{8} \cdot \frac{5}{9}$$

Como queremos calcular a probabilidade de as duas serem da mesma cor, elas poderão ser ambas brancas OU ambas vermelhas. Logo:

$$P\text{ (ambas da mesma cor)} = \frac{5}{8} \cdot \frac{4}{9} + \frac{3}{8} \cdot \frac{5}{9}$$

$$P\text{ (ambas da mesma cor)} = \frac{20}{72} + \frac{15}{72}$$

$$P\text{ (ambas da mesma cor)} = \frac{35}{72}$$

Nós somamos as duas probabilidades, porque, à operação lógica OU, associamos a operação aritmética soma.

Vamos agora calcular a probabilidade de as bolas serem de cores diferentes.

Duas situações podem ocorrer: a primeira bola ser branca e a segunda ser vermelha ou a primeira bola ser vermelha e a segunda ser branca.

$$P\text{ (1ª branca e 2ª vermelha)} = \frac{5}{8} \cdot \frac{5}{9}$$

$$P\text{ (1ª vermelha e 2ª branca)} = \frac{3}{8} \cdot \frac{4}{9}$$

Lembre-se de que só ocorre uma dessas duas situações: ou uma ou outra. Então:

$$P\text{ (bolas de cores diferentes)} = \frac{5}{8} \cdot \frac{5}{9} + \frac{3}{8} \cdot \frac{4}{9}$$

$$P\text{ (bolas de cores diferentes)} = \frac{25}{72} + \frac{12}{72} = \frac{37}{72}$$

Você pode, ainda, resolver a segunda parte desse exercício da seguinte forma: como eu sei que a probabilidade das bolas serem da mesma cor é $\frac{35}{72}$, então a probabilidade de serem de cores diferentes é o que falta para $\frac{72}{72}$, ou seja:

$$P\text{ (bolas de cores diferentes)} = 1 - \frac{35}{72} = \frac{37}{72}$$

4. Dois amigos foram caçar. Sabemos que um deles tem 45% de probabilidade de acertar qualquer caça, enquanto o outro tem 60%. Em cada tiro disparado, qual é a probabilidade de:

a) ambos acertarem na mesma caça;

b) nenhum acertar na mesma caça;

c) a caça ser atingida;

d) apenas um acertar a caça.

Vamos começar pela letra a.

A probabilidade de ambos acertarem a caça significa dizer que um E outro acertaram a caça. Então:

$$P\text{ (ambos acertarem a caça)} = \frac{45}{100} \cdot \frac{60}{100}$$

$$P\text{ (ambos acertarem a caça)} = \frac{27}{100}$$

Agora a letra b.

A probabilidade de nenhum acertar a caça significa que ambos erraram. Como um deles tem 45% de chance de acertar, deduzimos que tem 55% de chance de errar. O outro tem 60% de chance de acertar; logo tem 40% de chance de errar. Então:

$$P\text{ (nenhum acertar a caça)} = \frac{55}{100} \cdot \frac{40}{100}$$

$$P\text{ (nenhum acertar a caça)} = \frac{22}{100}$$

A letra c supõe que a caça foi atingida. Logo, sabemos que ou um, ou outro, ou ambos atingiram a caça. Não podemos garantir que ambos acertaram. Logo:

$$P\text{ (a caça ser atingida)} = \frac{45}{100} + \frac{60}{100} - \frac{45}{100} \cdot \frac{60}{100}$$

$$P\text{ (a caça ser atingida)} = \frac{105}{100} - \frac{27}{100}$$

$$P\text{ (a caça ser atingida)} = \frac{78}{100}$$

A letra d supõe que apenas um acertou a caça.

Quando a caça é atingida, ou foi atingida por apenas um, ou por ambos (como calculado na letra c), ou foi atingida por ambos (como calculado na letra a). Então:

$$P\text{ (apenas um acertar a caça)} = \frac{78}{100} - \frac{27}{100}$$

$$P\text{ (apenas um acertar a caça)} = \frac{51}{100}$$

Teorema de Bayes*

Ao estudarmos a probabilidade condicional, verificamos que $P(A \mid B)$ e $P(B \mid A)$ são diferentes.

Por exemplo, se A é o evento de uma pessoa ter sido aprovada em um concurso e se B é o evento de uma pessoa ter sido admitida em um emprego, $P(B|A)$ é a probabilidade de a pessoa ser admitida em um emprego, sabendo-se que, efetivamente, foi aprovada em um concurso. Já $P(A|B)$ é a probabilidade de a pessoa ter sido aprovada em um concurso, sabendo-se que, efetivamente, foi admitida em um emprego.

Vimos que:

$$P(A \cap B) = P(B) \cdot P(A \mid B) \text{ ou } P(A \cap B) = P(A) \cdot P(B \mid A)$$

Igualando essas duas expressões, temos:

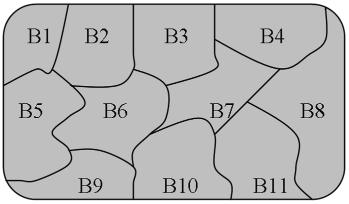
$$P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B)$$

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$$

Imaginemos agora um espaço amostral S, participando em um número finito de eventos B_i, onde i varia de 1 até n, de tal forma que a interseção entre quaisquer eventos B_i é um conjunto vazio (são eventos mutuamente exclusivos) e a união de todos os B_i é igual ao espaço amostral S.

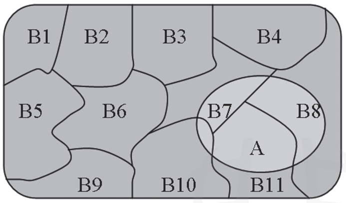
* Thomas Bayes (1702-1761) – reverendo e ministro presbiteriano.

Figura 1 – Partição de um espaço amostral S em um número finito de eventos B_i



Consideremos, agora, um evento A qualquer de S, em que P(A) > 0 (ver figura 2).

Figura 2 – Evento A pertencente ao espaço amostral S



Temos que:

$$A = (B_1 \cap A) \cup (B_2 \cap A) \cup \ldots \cup (B_n \cap A)$$

Então:

$$P(A) = P(B_1 \cap A) + P(B_2 \cap A) + \ldots + P(B_n \cap A)$$

Ou seja:

$$P(A) = \sum_{i=1}^{n} P(B_i \cap A)$$

Como $B_1, B_2, \ldots, B_n$ são eventos mutuamente exclusivos e como um evento B_i deve ocorrer, temos que:

$$P(B_K \mid A) = \frac{P(B_K) \cdot P(A \mid B_K)}{\sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i)}$$

Onde K = 1 ou 2 ou 3 ou ... ou n.

Dessa análise, concluimos que o teorema de Bayes é uma generalização de probabilidade condicional no caso de mais de dois eventos.[3]

Vamos aplicar essa sequência lógica do teorema no caso de uma fábrica de automóveis, em que três empresas de autopeças participam do processo de fabricação fornecendo peças dos tipos I, II e III, respectivamente. As peças do tipo I correspondem a 50%, as peças do tipo II correspondem a 20% e as peças do tipo III correspondem a 30% da produção total.

Considerando os dados do quadro abaixo, DETERMINE A PROBABILIDADE DE UMA PEÇA DEFEITUOSA QUE VENHA A SER UTILIZADA NA LINHA DE PRODUÇÃO SER PROVENIENTE DO FORNECEDOR DE PEÇAS DO TIPO III.

Fornecedores	Porcentagem de peças boas	Porcentagem de peças defeituosas
Fornecedores de peça I	98	2
Fornecedores de peça II	99	1
Fornecedores de peça III	97	3

Supondo:

A: o evento de uma peça defeituosa ser utilizada na linha de produção;

B_1: o evento da peça defeituosa ser do tipo I;

B_2: o evento da peça defeituosa ser do tipo II;

B_3: o evento da peça defeituosa ser do tipo III.

$$P(B_3 | A) = \frac{P(B_3) \cdot P(A | B_3)}{P(B_1) \cdot P(A | B_1) + P(B_2) \cdot P(A | B_2) + P(B_3) \cdot P(A | B_3)}$$

$$P(B_3 | A) = \frac{(0,30) \cdot (0,03)}{(0,50) \cdot (0,02) + (0,20) \cdot (0,01) + (0,30) \cdot (0,03)}$$

$$P(B_3 | A) = \frac{0,009}{0,021}$$

$$P(B_3 | A) = 0,4286 \text{ ou } 42,86\%$$

Síntese

Como afirmamos no início deste capítulo, "a probabilidade e a estatística estão estreitamente relacionadas, porque formulam tipos opostos de questões. Na probabilidade, sabemos como um processo ou experimento funciona e queremos predizer quais serão os resultados de tal processo. Em estatística, não sabemos como um processo funciona, mas podemos observar os seus resultados e utilizar as informações sobre os mesmos para conhecer a natureza do processo ou do experimento". Assim, trabalhando em direções diferenciadas, mas não isoladas, a conexão entre as duas áreas é bastante oportuna. No entanto, para que ocorra essa interligação, é necessário construir vários conhecimentos técnicos durante seu aprendizado. Por isso apresentamos aqui um panorama que se constituiu a partir das definições e conceitos básicos sobre a teoria das probabilidades, mas que revisitou também as concepções, usos e cálculos de componentes probabilísticos, como:

- esperança matemática;
- probabilidades finitas dos espaços amostrais finitos;
- espaços amostrais finitos equiprováveis;
- acontecimentos mutuamente exclusivos;
- regra da adição para eventos mutuamente exclusivos;
- eventos não mutuamente exclusivos;
- regra da adição para eventos não mutuamente exclusivos;
- acontecimentos simultâneos e sucessivos;
- probabilidade condicional;
- regra da multiplicação;
- independência estatística;
- Teorema de Bayes.

Todo esse processo, essa construção de conhecimentos, foi realizado por meio da contínua aplicação em contextos explicativos.

Questões para revisão

Nos exercícios a seguir, marque a alternativa correta.

1. Uma bola é retirada ao acaso de uma urna que contém seis bolas vermelhas, oito bolas pretas e quatro bolas verdes. Calcule a probabilidade de ela não ser preta.

   ( ) 10/18 ( ) 8/18

   ( ) 4/18 ( ) 12/18

   ( ) 6/18

2. A probabilidade de que Marcella resolva um problema é de 1/3 e a de que Luisa o resolva é de 1/4. Se ambas tentarem resolver independentemente o problema, qual a probabilidade de ele ser resolvido?

   ( ) 7/12 ( ) 2/7

   ( ) 1/7 ( ) 2/7

   ( ) 1/2

3. Jogamos, uma única vez, quatro moedas honestas. Qual a probabilidade, em qualquer ordem, de ter dado coroa em três das moedas e cara na outra moeda?

   ( ) 1/8 ( ) 3/16

   ( ) 3/8 ( ) 1/16

   ( ) 4/16

4. Uma carta é retirada de um baralho. Qual a probabilidade de ela ser uma dama ou uma carta de paus?

   ( ) 16/52 ( ) 4/52

   ( ) 17/52 ( ) 13/52

   ( ) 1/52

5. Uma empresa importadora tem 25% de chance de vender com sucesso um produto A e tem 40% de chance de vender com sucesso um produto B. Se essa empresa importar os dois produtos, A e B, qual a probabilidade de ela ter sucesso na venda ou do produto A, ou do produto B?

   ( ) 65/100 ( ) 75/100

   ( ) 55/100 ( ) 54/100

   ( ) 10/100

6. Uma empresa de vendas a varejo tem 60% de chance de vender o produto A com sucesso, enquanto o produto B tem apenas 40% de chance de ser vendido com sucesso. Se essa empresa tentar vender os produtos A e B, qual a probabilidade de ela ter sucesso na venda ou do produto A, ou do produto B?

   ( ) 100% ( ) 76%

   ( ) 60% ( ) 24%

   ( ) 40%

7. A ação da empresa A, negociada na Bolsa de Valores de São Paulo, tem 25% de chance de subir durante o pregão de determinado dia. Nesse mesmo dia, a ação da empresa B, negociada na mesma bolsa de valores, tem 20% de chance de subir. Caso um acionista compre ações das duas empresas, A e B, nesse dia, qual a probabilidade das ações de ambas subirem?

   ( ) 5% ( ) 22,5%

   ( ) 10% ( ) 40%

   ( ) 45%

8. Uma urna I contém quatro bolas vermelhas, três bolas pretas e três bolas verdes. Uma urna II contém duas bolas vermelhas, cinco bolas pretas e oito bolas verdes. Uma urna III contém dez bolas vermelhas, quatro bolas pretas e seis bolas verdes. Calcule a probabilidade de, retirando-se uma bola de cada urna, serem todas da mesma cor.

( ) 80/3000 ( ) 284/3000

( ) 60/3000 ( ) 140/3000

( ) 144/3000

9. Um pacote de sementes de flores contém quatro sementes de flores vermelhas, três de flores amarelas, duas de flores roxas e uma de flores de cor laranja. Escolhidas, ao acaso, uma após a outra, três sementes, qual a probabilidade de a primeira ser de flor de cor de laranja, a segunda de cor vermelha e a terceira de cor roxa?

( ) 7/27 ( ) 8/1000

( ) 242/720 ( ) 7/1000

( ) 8/720

10. Uma caixa contém vinte canetas iguais, das quais sete são defeituosas. Uma segunda caixa contém doze canetas iguais, das quais quatro são defeituosas. Considerando que uma caneta é retirada aleatoriamente de cada caixa, determine a probabilidade de uma ser perfeita e a outra não.

( ) 13/30 ( ) 11/20

( ) 9/20 ( ) 11/30

( ) 7/30

11. Uma pessoa tem dois automóveis velhos. Nas manhãs frias, há 20% de chance de um deles não pegar e 30% de chance de o outro não pegar. Qual a probabilidade de, em uma manhã fria, apenas um pegar?

( ) 24/100 ( ) 52/100

( ) 14/100 ( ) 38/100

( ) 50/100

12. Uma fábrica de louças tem um processo de inspeção com quatro etapas. A probabilidade de uma peça defeituosa passar numa etapa sem ser detectada é de aproximadamente 20%. Determine, então, a probabilidade de uma peça defeituosa passar por todas as quatro etapas de inspeção sem ser detectada.

( ) 0,20% ( ) 0,02%

( ) 0,0016% ( ) 0,80%

( ) 0,16%

13. Diferencie em que se baseia a estatística e em que se baseia o cálculo de probabilidades.

14. Quais as características dos fenômenos estudados em estatística?

capítulo 8

Distribuição binomial de probabilidades

Conteúdos do capítulo

- Aspectos da distribuição de probabilidades
- Aspectos da distribuição binomial
- Parâmetros da distribuição binomial
- Aplicação da distribuição binomial

Após o estudo deste capítulo, você será capaz de:

1. definir distribuição de probabilidades;
2. aplicar a distribuição binomial em casos práticos.

Distribuição de probabilidades

Na maioria dos problemas estatísticos, a amostra não é suficientemente grande para determinar a distribuição da população de maneira muito precisa. Contudo, há normalmente bastante informação na amostra, juntamente com a informação obtida de outras fontes, de modo a sugerir o tipo geral da distribuição da população correspondente. Combinando a experiência e a informação fornecidas pela amostra, podemos comumente convencionar a natureza geral da distribuição da população. Essa convenção leva ao que é conhecido como *distribuição de probabilidade* ou *distribuições teóricas*.[1]

> Uma distribuição de probabilidade é um modelo matemático para a distribuição real de frequências.

Em experimentos do tipo repetitivo (para os quais devemos construir um modelo de probabilidade) estamos geralmente interessados em uma propriedade particular da ocorrência do experimento. Por exemplo, no lançamento de três moedas, o interesse concentra-se sobre o número total de caras ou de coroas que aparece, porque é o que conta no jogo de moedas.

A quantidade escolhida para estudo em tal experimento será simbolizada pela letra x.

> Tratando-se de distribuição de probabilidade, a variável x é dita variável aleatória.

Com o objetivo de verificarmos como uma variável aleatória é introduzida num experimento simples, voltemos ao espaço amostral da experiência do lançamento de três moedas, já citado anteriormente no capítulo 7 e repetido a seguir.

Qual a probabilidade de obtermos uma única cara em um lançamento de três moedas?

Considerando K = cara, C = coroa e sendo A = {deu uma única cara}, temos:

S = {KCC, KKC, KKK, CKC, CKK, CCK, CCC, KCK}

Então, P(A) = 3/8.

Vamos analisar esse exemplo com mais detalhes.

Se x for usado para significar o número total de caras, então cada ponto do espaço amostral possuirá o valor x mostrado logo acima do ponto correspondente, como é representado a seguir.

3	2	2	2	1	1	1	0	⟵ variável x
.	.	.	.	.	.	.	.	
KKK	KKC	KCK	CKK	KCC	CKC	CCK	CCC	

Observemos que a variável x pode assumir qualquer dos valores 0, 1, 2 ou 3, mas nenhum outro valor. O valor da variável x só depende do ponto amostral particular escolhido. Isso significa que x é uma função dos pontos do espaço amostral.

Uma variável aleatória é aquela cujos valores são determinados por processos acidentais, ao acaso, que não estão sob o controle do observador. Ela pode ser denominada de *discreta* ou de *contínua*.

Para uma variável aleatória discreta, todos os possíveis valores da variável podem ser listados numa tabela com as probabilidades correspondentes. Os específicos modelos discretos de probabilidade que estudaremos são as distribuições de probabilidade binomial (neste capítulo) e a de Poisson (no próximo capítulo).

Para uma variável aleatória contínua, não podem ser listados todos os possíveis valores fracionários da variável e, dessa forma, as probabilidades determinadas por uma função matemática são retratadas, tipicamente, por uma função densidade ou por uma curva de probabilidade. Ela pode assumir qualquer valor numérico em um intervalo e, como exemplos, temos medidas como peso, altura e temperatura, entre outras. Nesse caso, estudaremos a distribuição normal de probabilidade (capítulo 10) e a distribuição qui-quadrado (capítulo 11).

Distribuição binomial

A distribuição binomial é uma distribuição discreta de probabilidade, aplicável sempre que o processo de amostragem é do tipo do de Bernoulli*. Um processo de Bernoulli é um processo de amostragem no qual:

---

* Daniel Bernoulli (1700-1782) – nasceu na Holanda, em Groningen, em uma família de grandes matemáticos. Obteve o bacharelado na Universidade de Basel, em 1715, e o mestrado em 1716.

a) em cada tentativa existem dois resultados possíveis e mutuamente exclusivos; eles são denominados, por conveniência, *sucesso* e *insucesso* (ou *fracasso*);

b) as séries de tentativas ou observações são constituídas de eventos independentes;

c) a probabilidade de sucesso, indicada por p, permanece constante de tentativa para tentativa; em outras palavras, o processo é estacionário.

Em geral, se p é a probabilidade de um evento acontecer em uma tentativa única, denominada *probabilidade de sucesso*, e $q = 1 - p$ é a probabilidade de que o evento não ocorra em qualquer tentativa única, denominada *probabilidade de insucesso*, então a probabilidade de o evento ocorrer exatamente X vezes em N tentativas, isto é, de que haja X sucessos e N – X insucessos, é dada por:

$$P(X) = C_{N,X} \cdot p^X \cdot q^{N-X} = \frac{N!}{X!\,(N-X)!} \cdot p^X \cdot q^{N-X}$$

O nome *binomial* se deve ao fato de que $C_{N,X} \cdot p^X \cdot q^{N-X}$ é o termo de ordem X em p no desenvolvimento do Binômio de Newton* $(q + p)^N$.

Nessa fórmula, precisamos recordar dois conceitos que você conhece da matemática: fatorial e combinação.

Mas não vamos perder tempo com teorias. Analisaremos alguns exemplos apenas para reativar e relembrar a prática.

Fatorial de um número N é representado por N! e é dado pela fórmula:

$N! = N \cdot (N-1) \cdot (N-2) \cdot (N-3) \cdot ... \cdot 1$

Então, vamos calcular o fatorial de 5.

$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$

$5! = 120$

Agora, vamos calcular o fatorial de 9.

$9! = 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$

$9! = 362880$

Como exemplo de variável binomial, temos o experimento que consiste no lançamento de uma moeda, no qual, em cada lance, podemos ter sucesso ou insucesso no resultado que desejamos.

* Isaac Newton (1642 - 1726) – notável matemático, físico e astrônomo inglês.

Para o cálculo das combinações de N elementos tomados X a X, representado por $C_{N,X}$, já mostrado anteriormente, basta utilizarmos a fórmula:

$$C_{N,X} = \frac{N!}{X!\,(N - X)!}$$

Agora, vamos calcular a combinação de cinco elementos tomados três a três. Veja.

$$C_{5,3} = \frac{5!}{3!\,(5 - 3)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1\,(2 \cdot 1)}$$

$$C_{5,3} = \frac{120}{6 \cdot 2}$$

$$C_{5,3} = 10$$

E, para terminar, vamos calcular a combinação de oito elementos tomados cinco a cinco.

$$C_{8,5} = \frac{8!}{5!\,(8 - 5)!}$$

$$C_{8,5} = \frac{40320}{120 \cdot 6}$$

$$C_{8,5} = 56$$

Aplicação da distribuição binomial

Agora, após recordar um pouco de matemática, vamos analisar exemplos da distribuição binomial de probabilidades.

No primeiro, vamos determinar a probabilidade de ocorrer três vezes o n° 6 em cinco (5) lances de um dado honesto.

$$P(X) = C_{5,3} \cdot p^3 \cdot q^{5-3}$$

> Lembre-se: ao jogar um dado uma única vez, a probabilidade de ocorrer qualquer resultado é igual a 1/6. Então, p = 1/6. Consequentemente, a probabilidade de não ocorrer esse resultado é igual a 5/6, pois a probabilidade de sucesso somada à probabilidade de insucesso é igual a 1 (ou seja, 100%). Então, q = 5/6.

Vamos, então, à solução dessa situação.

$$P(X) = \frac{5!}{3!\,(5-3)!} \cdot (1/6)^3 \cdot (5/6)^{5-3}$$

$$P(X) = 10 \cdot \frac{1}{216} \cdot \frac{25}{36}$$

$$P(X) = \frac{250}{7776}$$

P(X) = 0,03215 ou 3,215%

Exercício Resolvido

Verificou-se, em uma fábrica, que, em média, 10% dos parafusos produzidos por uma determinada máquina não satisfazem a certas especificações. Se forem selecionados, ao acaso, 10 parafusos da produção diária dessa máquina, DETERMINE A PROBABILIDADE DE EXATAMENTE TRÊS SEREM DEFEITUOSOS.

Atenção: sucesso é ocorrer o que desejamos. No caso, desejamos que os parafusos selecionados sejam defeituosos. Portanto, na estatística o sucesso não é necessariamente a parte boa de um experimento. Então, $p = 10\% = 0,1$ e $q = 90\% = 0,9$.

$P(3 \text{ defeituosos}) = C_{10,3} \cdot (0,1)^3 \cdot (0,9)^7$

P(3 defeituosos) = 120 · 0,001 · 0,4782969

P(3 defeituosos) = 0,0574 ou 5,74%.

Parâmetros da distribuição binomial

Supomos uma variável aleatória x que pode assumir o valor 1, no caso de sucesso, e o valor 0, no caso de insucesso (ou fracasso).

Sendo p a probabilidade de sucesso $P(x = 1) = p$ e q a probabilidade de insucesso $P(x = 0) = q$, temos que $p + q = 1$.

Havíamos definido N como o número de tentativas, e X como o número de sucessos. Então, X pode assumir valores de 1 até N e

$E(X) = E(X_1) + E(X_2) + \ldots + E(X_N) = N \cdot p$

Como as tentativas são independentes (distribuição de Bernoulli), a variância de X é igual à soma das variâncias de X_1, X_2, ..., X_N. Logo:

$S^2 = N \cdot p \cdot q$ e $S = \sqrt{N \cdot p \cdot q}$

Média: .... $\overline{X} = E(x) = N \cdot p$

Variância: .... $S^2 = N \cdot p \cdot q$

Desvio padrão: .... $S = \sqrt{S^2}$

Coeficiente de assimetria: .... $As = \dfrac{q - p}{S}$

Coeficiente de curtose: .... $K = 3 + \dfrac{1 - 6 \cdot p \cdot q}{S}$

Vamos aplicar esses cálculos?

Em 100 lances de uma moeda honesta, qual a média esperada de caras obtidas e qual o desvio padrão do experimento?

$\overline{X} = 100 \cdot (1/2) = 50$

$S^2 = 100 \cdot (1/2) \cdot (1/2) = 25$

Então: $S = 5$

PORTANTO, A MÉDIA É DE 50 E O DESVIO PADRÃO É DE 5.

Observe novamente que, para a determinação de p e de q, supomos um único lançamento da moeda. Lembrar que a moeda é honesta. Logo, a probabilidade de dar cara é a mesma probabilidade de dar coroa.

Exercícios Resolvidos

1. Devido às altas taxas de juros, uma empresa informa que 30% de suas contas a receber de outras empresas comerciais encontram-se vencidas. Se um contador escolher, aleatoriamente, uma amostra de cinco contas, DETERMINE A PROBABILIDADE DE EXATAMENTE 20% DESSAS CONTAS ESTAREM VENCIDAS.

Importante lembrar que 20% das contas escolhidas é:

20% de 5 = $0{,}20 \cdot 5 = 1$ (ou seja, queremos saber a chance de uma única conta estar vencida entre as cinco que foram escolhidas pelo contador).

$$P(X) = C_{N,X} \cdot p^X \cdot q^{N-X} = \frac{N!}{X!\,(N-X)!} \cdot p^X \cdot q^{N-X}$$

Lembre-se, também, que 30% das contas estão vencidas.

Como estamos tentando encontrar uma conta vencida, a probabilidade de sucesso é p = 30%, ou seja, p = 0,3. Logo, q = 0,7.

Então:

$$P(X = 1) = C_{5,1} \cdot 0{,}3^1 \cdot 0{,}7^{5-1}$$

$$P(X = 1) = \frac{5!}{1!\,(5-1)!} \cdot 0{,}3 \cdot 0{,}2401$$

$$P(X=1) = 5 \cdot 0{,}3 \cdot 0{,}2401$$

$$P(X=1) = 0{,}36015 \text{ ou } 36{,}015\%$$

2. Uma firma de pedidos pelos correios enviou uma carta circular que tem uma taxa de respostas de 10%. Suponha que 20 cartas circulares são endereçadas a uma nova área geográfica, como um teste de mercado. Considerando que na nova área é aplicável a taxa de respostas de 10%, DETERMINE A PROBABILIDADE, USANDO A FÓRMULA DE PROBABILIDADES BINOMIAIS, DE APENAS UMA PESSOA RESPONDER.

Como a probabilidade de resposta é de 10%, e o sucesso (o que queremos que aconteça) é que uma pessoa responda, p = 0,1, logo, q = 0,9.

A minha amostra é de 20 cartas. Então, N = 20.

Como queremos determinar a probabilidade de uma pessoa responder, X = 1.

$$P(X) = C_{N,X} \cdot p^X \cdot q^{N-X} = \frac{N!}{X!\,(N-X)!} \cdot p^X \cdot q^{N-X}$$

$$P(X) = C_{20,1} \cdot 0{,}1^1 \cdot 0{,}9^{20-1}$$

$$P(X = 1) = \frac{20!}{1!\,(20-1)!} \cdot 0{,}1 \cdot 0{,}135085171$$

$$P(X=1) = 20 \cdot 0{,}1 \cdot 0{,}135085171$$

$$P(X=1) = 0{,}27017 \text{ ou } 27{,}017\%$$

3. Durante um ano particular, 70% das ações ordinárias negociadas na Bolsa de Valores do Rio de Janeiro tiveram sua cotação aumentada, enquanto 30% tiveram sua cotação diminuída ou estável. No começo do ano, um serviço de assessoria financeira escolhe dez ações como sendo especialmente recomendadas. Se as dez ações representam uma seleção aleatória, usando a fórmula de probabilidades binomiais, qual a probabilidade de que todas as dez ações escolhidas tenham tido suas cotações aumentadas?

Escolhemos dez ações. Então, N = 10.

Desejamos que as dez tenham tido sucesso, ou seja, as dez aumentaram suas cotações. Então, X = 10.

A probabilidade de sucesso é de 70%, logo p = 0,7, em consequência, q = 0,3.

$$P(X) = C_{N,X} \cdot p^X \cdot q^{N-X} = \frac{N!}{X!\,(N-X)!} \cdot p^X \cdot q^{N-X}$$

$$P(X=10) = C_{10,10} \cdot 0{,}7^{10} \cdot 0{,}3^{10-10}$$

$$P(X=10) = \frac{10!}{10!\,(10-10)!} \cdot 0{,}028248 \cdot 1$$

Lembre-se: o fatorial de zero, por definição, é igual a 1.

Lembre-se, também, de que qualquer número real elevado a zero é igual a 1.

$$P(X=10) = 1 \cdot 0{,}028248 \cdot 1$$

$$P(X=10) = 0{,}028248 \text{ ou } 2{,}8248\%$$

Pelo resultado, verificamos que a chance de as dez ações escolhidas terem, todas elas, as suas cotações aumentadas é muito pequena.

Síntese

Neste capítulo, o objetivo foi definir a distribuição de probabilidades e construir com você, por meio de uma série de aplicações, a competência na aplicação da distribuição binomial. Para isso, foi necessário fazer vários cálculos, descrevendo situações diversas de aplicação, e verificar os parâmetros da distribuição binomial com suas respectivas equações matemáticas: média, variância, desvio padrão, coeficiente de assimetria e coeficiente de curtose.

Questões para revisão

Nos exercícios a seguir, assinale a alternativa correta.

1. Verifica-se, em uma fábrica, que, em média, 10% dos parafusos produzidos por uma determinada máquina não satisfazem a certas especificações. Se forem selecionados, ao acaso, oito parafusos da produção diária dessa máquina, usando a fórmula de probabilidades binomiais, determine a probabilidade de nenhum deles ser defeituoso.

   ( ) 0,431%   ( ) 43,05%   ( ) 64,98%

   ( ) 4,305%   ( ) 6,498%

2. Em um concurso realizado para trabalhar em determinada empresa de exportação, 10% dos candidatos foram aprovados. Se escolhermos, aleatoriamente, dez candidatos desse concurso, qual a probabilidade de que exatamente dois deles tenham sido aprovados?

   ( ) 0,43%   ( ) 43%   ( ) 19,37%

   ( ) 4,3%   ( ) 0,1937%

3. Em determinada turma de uma universidade, em 2006, 20% dos alunos foram reprovados em Matemática Comercial e Financeira. Se escolhermos, aleatoriamente, oito alunos dessa turma, qual a probabilidade de exatamente três desses alunos terem sido reprovados?

   ( ) 32,77%   ( ) 16,39%   ( ) 7,32%

   ( ) 0,8%   ( ) 14,68%

4. Qual a probabilidade de obtermos exatamente cinco coroas em seis lances de uma moeda não viciada?

   ( ) 9,375%   ( ) 15,625%   ( ) 4,375%

   ( ) 1,5625%   ( ) 10,9375%

5. Em um ano particular, 30% dos alunos de determinada faculdade de medicina do Estado de São Paulo foram reprovados em Clínica Geral. Se escolhermos, aleatoriamente, dez alunos dessa universidade que tenham cursado Clínica Geral, qual a probabilidade de exatamente três deles terem sido reprovados?

( ) 14,68%   ( ) 26,68%   ( ) 10,94%

( ) 2,7%   ( ) 8,24%

6. Defina o que é uma distribuição de probabilidades.
7. O que é uma variável aleatória?

capítulo 9

Distribuição de Poisson de probabilidades

Conteúdos do capítulo

- Conceituação da distribuição de Poisson
- Aplicação da distribuição de Poisson
- Parâmetros da distribuição de Poisson

Após o estudo deste capítulo, você será capaz de:

1. definir a distribuição de Poisson;
2. aplicar em casos práticos a distribuição de Poisson.

Distribuição de Poisson*

A distribuição de Poisson pode ser usada para determinar a probabilidade de um dado número de sucessos quando os eventos ocorrem em um *continuum*** de tempo ou espaço. Tal processo, chamado de *processo de Poisson*, é similar ao processo de Bernoulli, exceto que os eventos ocorrem em um *continuum* em vez de de ocorrerem em tentativas ou observações fixadas. Um exemplo de tal processo é a chegada de chamadas telefônicas em uma central telefônica. Tal como no caso de Bernoulli, supomos que os eventos são independentes e que o processo é estacionário.

Somente um valor é necessário para determinar a probabilidade de um dado número de sucessos em um processo de Poisson: o número médio de sucessos para a específica dimensão de interesse. Esse número médio é geralmente representado por λ (lambda)*** ou por μ (mu). A fórmula para determinarmos a probabilidade de um dado número X de sucessos em uma distribuição de Poisson é:

$$P(X \mid \lambda) = \frac{\lambda^X \cdot e^{-\lambda}}{X!}$$

em que "e" é a constante 2,71828 e é a base do logaritmo natural (neperiano), sendo que os valores de $e^{-\lambda}$ podem ser obtidos na Tabela 34 (a seguir). A representação da $P(X \mid \lambda)$, lê-se a probabilidade de acontecer X tal que a média conhecida é λ.

---

* Denis Poisson – foi um matemático francês (1781-1840).

** Série longa de elementos, em uma determinada sequência, em que cada um difere minimamente do elemento subsequente, daí resultando diferença acentuada entre os elementos iniciais e finais da sequência (Dicionário Eletrônico Houaiss da Língua Portuguesa).

*** λ é a letra grega minúscula que lemos "lambda" e que corresponde à letra *l* do nosso alfabeto.

Caso não tenhamos acesso a uma tabela com os valores de $e^{-\lambda}$, é necessário dispor de uma calculadora científica que nos permita determinar tais valores.

Até agora, temos representado a média ou média aritmética por $\overline{X}$. No entanto, para o estudo das distribuições de Poisson e normal, nós a representaremos por λ.

Tabela 34 – Valores de $e^{-\lambda}$ para alguns valores de λ.

λ	$e^{-\lambda}$	λ	$e^{-\lambda}$
0,0	1,00000	2,5	0,08208
0,1	0,90484	2,6	0,07427
0,2	0,81873	2,7	0,06721
0,3	0,74082	2,8	0,06081
0,4	0,67032	2,9	0,05502
0,5	0,60653	3,0	0,04979
0,6	0,54881	3,2	0,04076
0,7	0,49659	3,4	0,03337
0,8	0,44933	3,6	0,02732
0,9	0,40657	3,8	0,02237
1,0	0,36788	4,0	0,01832
1,1	0,33287	4,2	0,01500
1,2	0,30119	4,4	0,01228
1,3	0,27253	4,6	0,01005
1,4	0,24660	4,8	0,00823
1,5	0,22313	5,0	0,00674
1,6	0,20190	5,5	0,00409
1,7	0,18268	6,0	0,00248
1,8	0,16530	6,5	0,00150
1,9	0,14957	7,0	0,00091
2,0	0,13534	7,5	0,00055
2,1	0,12246	8,0	0,00034
2,2	0,11080	8,5	0,00020
2,3	0,10026	9,0	0,00012
2,4	0,09072	10,0	0,00005

Aplicação da distribuição de Poisson

Para entendermos bem esse assunto, nada melhor do que resolvermos alguns casos:

Por exemplo, se um departamento de conserto de máquinas recebe uma média de cinco chamadas por hora, QUAL A PROBABILIDADE DE QUE, EM UMA HORA, SELECIONADAS ALEATORIAMENTE, SEJAM RECEBIDAS EXATAMENTE TRÊS CHAMADAS?

$$P(X = 3 \mid \lambda = 5) = \frac{5^3 \cdot e^{-5}}{3!}$$

Lembre-se de verificar na tabela 34 o valor de e^{-5}:

$$P(X = 3 \mid \lambda = 5) = \frac{125 \cdot 0{,}00674}{6} = 0{,}1404 \text{ ou } 14{,}04\%$$

PORTANTO, A PROBABILIDADE É DE 14,04%.

Já, se o caso for uma experiência de laboratório em que passam, em média, em um contador de partículas, quatro partículas radioativas por milissegundo, QUAL A PROBABILIDADE DE ENTRAREM NO CONTADOR SEIS PARTÍCULAS RADIOATIVAS EM DETERMINADO MILISSEGUNDO?

A média conhecida é igual a 4, ou seja: $\lambda = 4$

$$P(X = 6 \mid \lambda = 4) = \frac{\lambda^X \cdot e^{-\lambda}}{X!}$$

$$P(X = 6 \mid \lambda = 4) = \frac{4^6 \cdot e^{-4}}{6!}$$

$$P(X = 6 \mid \lambda = 4) = \frac{4096 \cdot 0{,}01832}{720}$$

$$P(X = 6 \mid \lambda = 4) = 0{,}10422 \text{ ou } 10{,}422\%$$

A PROBABILIDADE É DE 10,422%.

E, considerando que nos sinais de um transmissor ocorrem distorções aleatórias a uma taxa média de uma por minuto, e as mensagens têm, em média, três (3) minutos, ou seja, a ocorrência de distorção é de três por mensagem, QUAL A PROBABILIDADE DE O NÚMERO DE DISTORÇÕES EM UMA MENSAGEM DE TRÊS MINUTOS SER IGUAL A DOIS (2), UTILIZANDO A FÓRMULA DE POISSON ?

Como, em média, ocorre uma distorção a cada minuto, em três minutos (tempo da mensagem) deverão ocorrer três distorções. Então, $\lambda = 3$.

$$P(X = 2 \mid \lambda = 3) = \frac{\lambda^X \cdot e^{-\lambda}}{X!}$$

$$P(X = 2 \mid \lambda = 3) = \frac{3^2 \cdot e^{-3}}{2!}$$

$$P(X = 2 \mid \lambda = 3) = \frac{9 \cdot 0{,}04979}{2}$$

$$P(X = 2 \mid \lambda = 3) = 0{,}2241 \text{ ou } 22{,}41\%$$

Logo, a probabilidade é de 22,41%.

No caso de, em média, quatro pessoas utilizarem, por hora, um determinado telefone público, e sabendo que o aparelho, quando apresenta defeito, é reparado após uma hora da comunicação do defeito, qual a probabilidade de que somente o reclamante (supondo que ele foi o primeiro a constatar o defeito) não possa usar o telefone devido àquele defeito?

A média conhecida é $\lambda = 4$ (quatro pessoas por hora utilizam determinado telefone público). Queremos conhecer a probabilidade de apenas uma pessoa utilizar tal telefone em determinada hora ($X = 1$).

$$P(X = 1 \mid \lambda = 4) = \frac{\lambda^X \cdot e^{-\lambda}}{X!}$$

$$P(X = 1 \mid \lambda = 4) = \frac{4^1 \cdot e^{-4}}{1!}$$

$$P(X = 1 \mid \lambda = 4) = \frac{4 \cdot 0{,}01832}{1}$$

$$P(X = 1 \mid \lambda = 4) = 0{,}07328 \text{ ou } 7{,}328\%$$

Aqui a probabilidade é de 7,328%.

Considere que, em média, um digitador comete três erros a cada 6.000 números teclados. Qual a probabilidade de que, na digitação de um importante relatório, composto por 2.000 números, não ocorram erros?

Como são cometidos, em média, três erros a cada 6.000 números teclados, para 2.000 números teclados é esperado apenas um erro (aplicando-se a regra de três simples).

Então, $\lambda = 1$ (para 2.000 números teclados).

Queremos determinar a probabilidade de não ocorrerem erros, ou seja, $X = 0$.

$$P(X = 0 \mid \lambda = 1) = \frac{1^0 \cdot e^{-1}}{0!}$$

$$P(X = 0 \mid \lambda = 1) = \frac{1 \cdot 0{,}36788}{1}$$

$$P(X = 0 \mid \lambda = 1) = 0{,}36788 \text{ ou } 36{,}788\%$$

Portanto, a probabilidade de erros é de 36,788%.

Exercício Resolvido

1. Em cada dez dias chegam, em média, trinta navios à determinada doca. Qual a probabilidade de que, em um dia, aleatoriamente escolhido, cheguem à doca exatamente quatro navios?

Como em dez dias chegam, em média, 30 navios, em um dia espera-se que cheguem três navios (por regra de três simples), ou seja: $\lambda = 3$.

$$P(X = 4 \mid \lambda = 3) = \frac{3^4 \cdot e^{-3}}{4!}$$

$$P(X = 4 \mid \lambda = 3) = \frac{81 \cdot 0{,}04979}{24}$$

$$P(X = 4 \mid \lambda = 3) = 0{,}1680 \text{ ou } 16{,}80\%$$

Resposta: A probabilidade da chegada de quatro navios é de 16,80%.

Parâmetros da distribuição de Poisson

Observe que, quando uma variável aleatória x admite distribuição binomial com o número N de tentativas muito grande ($N > 30$) e com a probabilidade p de sucesso muito pequena, o uso da fórmula $P(X) = C_{N,X} \cdot p^X \cdot q^{N-X}$ torna-se praticamente inviável[1].

Nessas condições, o valor de P(X) é aproximadamente igual a $\frac{e^{-\lambda} \cdot \lambda^X}{X!}$, em que $\lambda = N \cdot p$ (mesma média da distribuição binomial).

Média: .... $\lambda = N \cdot p$

Variância: .... $S^2 = N \cdot p \cdot q$

Desvio padrão: .... $S = \sqrt{S^2}$

Coeficiente de assimetria: .... $As = 1/S$

Coeficiente de curtose: .... $K = 3 + 1/S^2$

Vamos aplicar esses parâmetros?

Sendo que a probabilidade de um indivíduo sofrer uma reação alérgica, resultante da injeção de determinado soro, é igual a 0,001, determine, então, a probabilidade de, entre 2.000 indivíduos, exatamente três sofrerem a mesma reação alérgica.

Seja X = {indivíduo sofre reação alérgica}

$$P(X \mid \lambda) = \frac{\lambda^X \cdot e^{-\lambda}}{X!}$$

Observe que a média não foi fornecida. Portanto, para aplicarmos Poisson, precisamos calculá-la.

Temos que $\lambda = N \cdot p$

$\lambda = 2.000 \cdot 0{,}001$

$\lambda = 2$

Logo:

$$P(X = 3 \mid \lambda = 2) = \frac{2^3 \cdot e^{-2}}{3!}$$

$$P(X = 3 \mid \lambda = 2) = \frac{8 \cdot 0{,}13534}{6}$$

$P(X = 3 \mid \lambda = 2) = 0{,}1805$ ou 18,05%

Logo a probabilidade é de 18,05%.

Exercícios Resolvidos

1. A probabilidade da ocorrência de morte por Aids, entre os habitantes de determinada comunidade, é de 0,002 por ano. Uma empresa de seguros forma um contrato coletivo com essa comunidade que tem 5.000 habitantes. Nesse contexto, qual a probabilidade de que cinco habitantes desse grupo morram de Aids no primeiro ano de existência da apólice?

A média de mortalidade não foi fornecida. Precisamos calculá-la.

$\lambda = N \cdot p$

$\lambda = 5.000 \cdot 0{,}002$

$\lambda = 10$

Logo:

$$P(X = 5 \mid \lambda = 10) = \frac{10^5 \cdot e^{-10}}{5!}$$

$$P(X = 5 \mid \lambda = 10) = \frac{100.000 \cdot 0{,}00005}{120}$$

$P(X = 5 \mid \lambda = 10) = 0{,}04167$ ou 4,167%

Resposta: A probabilidade é de 4,167%.

2. Uma companhia de seguros está considerando a inclusão da cobertura de uma doença relativamente rara na área geral de seguros médicos. A probabilidade de que um indivíduo, selecionado aleatoriamente, venha a contrair a doença é 0,001, sendo que 3.000 pessoas estão incluídas no grupo segurado. Sendo assim, qual o número esperado de pessoas, no grupo, que terão a doença? Qual a probabilidade de que nenhuma das 3.000 pessoas do grupo contraia a doença?

A média esperada (número esperado de pessoas que terão a doença) é:

$\lambda = N \cdot p$

$\lambda = 3000 \cdot 0{,}001$

$\lambda = 3$

Significa que o número esperado de pessoas que terão a doença é de 3.

Logo:

$$P(X = 0 \mid \lambda = 3) = \frac{3^0 \cdot e^{-3}}{0!}$$

$$P(X = 0 \mid \lambda = 3) = \frac{1 \cdot 0{,}04979}{1}$$

$$P(X = 0 \mid \lambda = 3) = 0{,}04979 \text{ ou } 4{,}979\%$$

Resposta: A probabilidade de nenhuma contrair a doença é de 4,979%.

3. A probabilidade de uma pessoa sofrer intoxicação alimentar na lanchonete de determinado bairro de baixa renda é de 0,001. Determine a probabilidade de que, das 1.000 pessoas que vão por dia nessa lanchonete, exatamente duas se intoxiquem.

A média esperada de intoxicação é:

$$\lambda = N \cdot p$$

$$\lambda = 1000 \cdot 0{,}001$$

$$\lambda = 1$$

Logo:

$$P(X = 2 \mid \lambda = 1) = \frac{1^2 \cdot e^{-1}}{2!}$$

$$P(X = 2 \mid \lambda = 1) = \frac{1 \cdot 0{,}36788}{2}$$

$$P(X = 2 \mid \lambda = 1) = 0{,}18394 \text{ ou } 18{,}394\%$$

Resposta: Portanto, a probabilidade é de 18,394%.

Síntese

Neste capítulo, tratamos especificamente da definição e aplicação da distribuição de Poisson. Tal distribuição, elaborada pelo matemático francês Denis Poisson, serve para verificarmos as probabilidades em eventos que ocorrem em movimento contínuo, ao contrário da distribuição binomial (desenvolvida por Bernoulli), que ocorrem em observações fixas. Os parâmetros referem-se aos mesmos fenômenos, em ambas, isto é: média, variância, desvio padrão, coeficiente de assimetria e coeficiente de curtose.

Questões para revisão

Assinale a alternativa correta nos exercícios abaixo.

1. Na fabricação de resistores de 50 ohms, são considerados bons os que têm resistência entre 45 e 55 ohms. Sabe-se que a probabilidade de um deles ser defeituoso é 0,2%, sendo que são vendidos em lotes de 1.000 unidades. Nesse caso, qual a probabilidade de um resistor ser defeituoso, em um lote?

   ( ) 13,534% ( ) 27,068% ( ) 0,271%

   ( ) 6,767% ( ) 0,135%

2. Sendo a probabilidade de uma pessoa sofrer reação alérgica, resultante da injeção de determinado soro, ser igual a 0,0002, determine a probabilidade de, entre 5.000 pessoas, exatamente três sofrerem a mesma reação alérgica.

   ( ) 36,788% ( ) 13,534% ( ) 0,674%

   ( ) 0,833% ( ) 6,13%

3. Em média, oito pessoas por dia consultam um especialista em decoração de determinada fábrica. Qual a probabilidade de que, em um dia selecionado aleatoriamente, exatamente três pessoas façam tal consulta?

( ) 2,90% ( ) 14,66% ( ) 58,02%

( ) 29,01% ( ) 5,80%

4. Um departamento de conserto de máquinas recebe, em média, quatro chamadas por hora. Qual a probabilidade de que, em uma hora, selecionada aleatoriamente, sejam recebidas exatamente duas chamadas?

   ( ) 1,83% ( ) 7,33% ( ) 18,30%

   ( ) 14,66% ( ) 3,66%

5. Em Tóquio, ocorrem, em média, seis suicídios por mês. Calcule a probabilidade de, em um mês, selecionado aleatoriamente, ocorrer exatamente dois suicídios.

   ( ) 0,74% ( ) 8,93% ( ) 4,46%

   ( ) 7,44% ( ) 44,64%

6. Defina a utilização da distribuição de Poisson.

7. Quantos valores são necessários para determinar a probabilidade de um dado número de sucessos em um processo de Poisson?

capítulo 10

Distribuição normal de probabilidades

Conteúdos do capítulo

- Conceituação e características da distribuição normal
- Aplicação da distribuição normal
- Parâmetros da distribuição normal

Após o estudo deste capítulo, você será capaz de:

1. definir distribuição normal de probabilidades;
2. aplicar a distribuição normal de probabilidades em situações concretas.

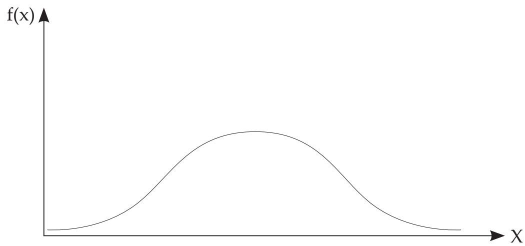
Distribuição normal

A distribuição normal de probabilidade é uma distribuição de probabilidade contínua que é simétrica em relação à média e mesocúrtica e assíntota em relação ao eixo das abscissas, em ambas as direções. A curva que representa a distribuição normal de probabilidade é frequentemente descrita como tendo uma forma de sino, sendo também conhecida como *Curva de Gauss** (veja a figura 3).

Inicialmente, supunha-se que todos os fenômenos da vida real devessem ajustar-se a uma curva em forma de sino; caso contrário, suspeitava-se de alguma anormalidade no processo de coleta de dados. Daí a designação de *curva normal*[1].

* Joham Carl Friedrich Gauss (1777-1855) – famoso matemático, astrônomo e físico alemão, era conhecido como o *príncipe dos matemáticos*.

Figura 3 – Curva representativa da distribuição normal de probabilidade



A distribuição de probabilidade normal é importante na inferência estatística por três razões distintas:

a) as medidas produzidas em diversos processos aleatórios seguem essa distribuição;

b) as probabilidades normais podem ser usadas frequentemente como aproximações de outras distribuições de probabilidade, tais como a binomial e a de Poisson;

c) as distribuições de estatísticas da amostra, tais como a média e a proporção, frequentemente seguem a distribuição normal independentemente da distribuição da população.

Como para qualquer distribuição contínua de probabilidade, o valor da probabilidade pode somente ser determinado para um intervalo de valores da variável.

A altura da função densidade (ou curva de probabilidade) para uma variável normalmente distribuída é dada por:

$$f(X) = \frac{1}{S \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2}\left(\frac{X-\lambda}{S}\right)^2}$$

Em que:

$f(X)$ = variável dependente.

X = variável independente.

λ = média.

S = desvio padrão da distribuição.

e = 2,71828... (base do sistema de logaritmos neperianos).

π = 3,14159...

Observe que, mais uma vez, representamos a média (ou média aritmética) pela letra λ.

Uma vez que cada combinação de λ e de S geraria uma distribuição normal de probabilidade diferente (todas simétricas e mesocúrticas), as tabelas de probabilidades da normal são baseadas em uma distribuição normal de probabilidade com $\lambda = 0$ e $S = 1$.

Qualquer conjunto de valores X, normalmente distribuídos, pode ser convertido em valores normais z padronizados pelo uso da fórmula:

$$z = \frac{X - \lambda}{S}$$

A Tabela 35 indica as proporções de área para vários intervalos de valores para a distribuição de probabilidade normal padronizada, com a fronteira inferior do intervalo começando sempre na média. A conversão dos valores dados da variável X em valores padronizados torna possível o uso dessa tabela e faz com que seja desnecessário o uso da equação de densidade de qualquer distribuição normal dada.

Tabela 35 – Áreas de uma distribuição normal padrão*.

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3448	0,3461	0,3485	0,3508	0,3531	0,3534	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4396	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4703
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4922	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

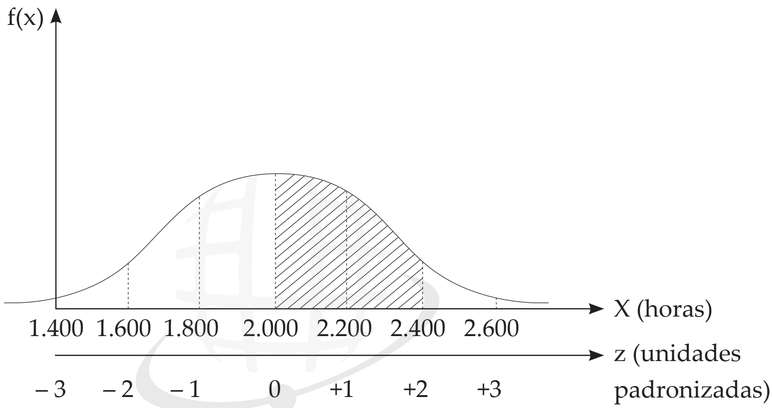
* Cada casa na tabela dá a proporção sob a curva inteira entre z = 0 e um valor positivo de z. As áreas para os valores de z negativos são obtidas por simetria.

Aplicação da distribuição normal

Vamos, agora, conferir, por meio de um exemplo prático, a distribuição de probabilidade padronizada?

Sabemos que a vida útil de determinado componente elétrico segue uma distribuição normal com média λ = 2.000 horas e desvio padrão S = 200 horas. Qual é a probabilidade de que um componente elétrico, aleatoriamente selecionado, dure entre 2.000 e 2.400 horas?

Figura 4 – Curva de probabilidade para vida útil do componente



A Figura 4 retrata a curva de probabilidade (função densidade) para esse problema e indica, também, a relação entre a escala de horas X e a escala da normal padronizada z.

Observe a área sob a curva correspondente ao intervalo de 2.000 a 2.400 horas.

A fronteira inferior do intervalo está na média da distribuição e, portanto, está no valor z = 0. A fronteira superior do intervalo, em termos de valor de z, é:

$$z = \frac{X - \lambda}{S} = \frac{2.400 - 2.000}{200} = +2,0$$

De acordo com a Tabela 35, verificamos que:

$P(0 \leq z \leq +2,0) = 0,4772$

Portanto:

$P(2.000 \leq X \leq 2.400) = 0,4772$ ou 47,72 %.

(Expectativa de vida útil do componente.)

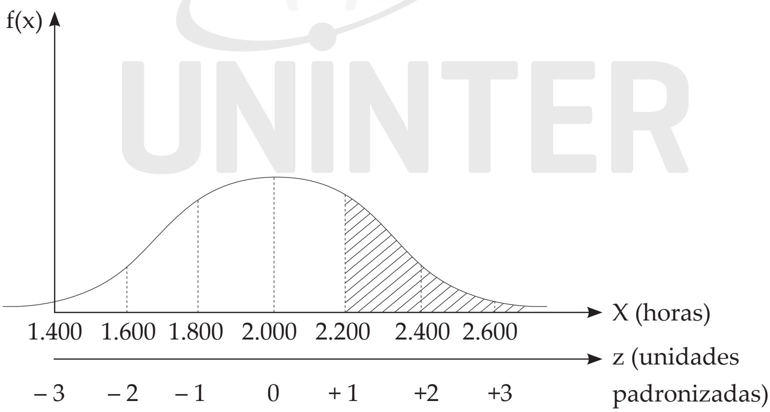
O que isso significa? Significa que a área limitada pela curva e pelo eixo horizontal X, para valores de z, variando de 0 até +2, corresponde a 47,72% da área total.

É claro que nem todos os problemas envolvem intervalos tendo a média como fronteira inferior. Contudo, a tabela de áreas de uma distribuição normal padrão pode ser usada para determinar o valor de probabilidade associado com qualquer intervalo dado, tanto por adição ou subtração apropriadas de áreas como pelo reconhecimento de que a curva é simétrica.

Agora, com respeito aos componentes elétricos descritos no caso analisado, suponhamos uma segunda situação: que estivéssemos interessados na probabilidade de que um componente, aleatoriamente escolhido, durasse mais do que 2.200 horas.

Observe que o ponto em que z = 0 corresponde à metade da curva normal. Logo a proporção total da área à direita da média na Figura 5 é igual a 0,5000, 50% da área total. Portanto, se determinarmos a proporção entre a média e 2.200, podemos subtrair esse valor de 0,5 para obtemos a probabilidade de que as horas X sejam maiores do que 2.200.

Figura 5 – Curva de probabilidade para componente com mais de 2200 horas de vida útil



$$z = \frac{2.200 - 2.000}{200} = +1,0$$

$P(0 \leq z \leq +1,0) = 0,3413$

$P(z > +1,0) = 0,5000 - 0,3413 = 0,1587$

Portanto:

$P(X > 2200) = 0,1587$ ou 15,87%.

Assim, a probabilidade do componente durar mais de 2.200 horas é de 15,87%.

Parâmetros da distribuição normal

Média: .... $\lambda = N \cdot p$

Variância: .... $S^2 = N \cdot p \cdot q$

Desvio padrão: .... $S = \sqrt{S^2}$

Coeficiente de assimetria: .... $As = 0$

Coeficiente de curtose: .... $K = 3$

Vamos aplicar esses parâmetros em situações concretas para melhor entendimento. Nesse contexto, considerando que os pesos dos alunos de determinada escola têm uma distribuição normal com média de 50 kg e desvio padrão de 5 kg, vamos verificar qual a porcentagem de alunos dessa escola com peso:

a) abaixo de 45 kg;

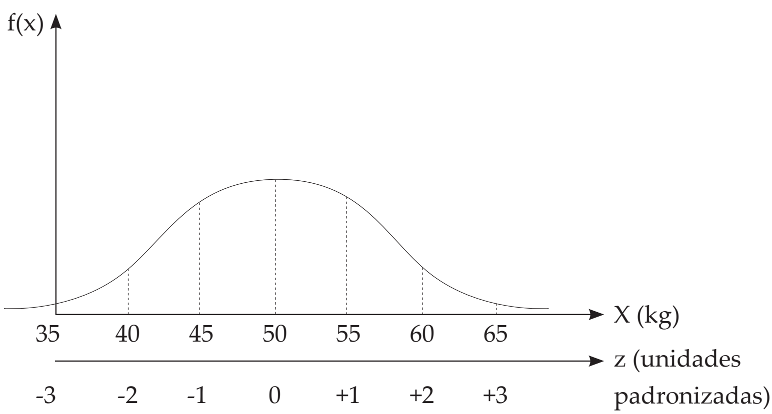
b) acima de 60 kg;

c) entre 48 kg e 58 kg.

Vamos começar calculando quantos alunos têm peso abaixo de 45 kg.

Para melhor visualizarmos o problema, é sempre recomendável que façamos o gráfico correspondente (veja a Figura 6).

Figura 6 – Curva de probabilidade para porcentagem de alunos x peso



Como estamos interessados em todos os alunos abaixo de 45 kg, desejamos saber qual é a área abaixo da curva e à esquerda de 45 kg.

Então, para X = 45, temos:

$$z = \frac{X - \lambda}{S} = \frac{45 - 50}{5} = -1,0$$

$P(-1 \leq z \leq 0) = 0,3413$

$P(z < -1,0) = 0,5000 - 0,3413 = 0,1587$

Portanto:

$P(X < 45) = 0,1587$ ou 15,87 %.

Vamos calcular agora que percentual de alunos tem acima de 60 kg, ou seja, $X \geq 60$.

$$z = \frac{X - \lambda}{S} = \frac{60 - 50}{5} = +2,0$$

$P(0 \leq z \leq +2,0) = 0,4772$

Portanto:

$P(50 \leq X \leq 60) = 0,4772$ ou 47,72 %

Então, acima de 60 kg temos:

$P(X > 60) = 50\ \% - 47,72\ \% = 2,28\ \%$

Por último, vamos verificar quantos alunos pesam entre 48 kg e 58 kg.

Para X = 48, temos:

$$z = \frac{X - \lambda}{S} = \frac{48 - 50}{5} = -0,4$$

Para X = 58, temos:

$$z = \frac{X - \lambda}{S} = \frac{58 - 50}{5} = +1,6$$

> Agora lembre-se: toda análise deve ser feita a partir da média, ou seja, a partir de z igual a zero.

Estamos interessados em saber o percentual abaixo da curva entre os valores de:

$z = -0,4$ e $z = +1,6$.

Então:

$P(-0,4 \leq z \leq 0) = 0,1554$

Portanto:

$P(48 \leq X \leq 50) = 0,1554$ ou 15,54%

e $P(0 \leq z \leq 1,6) = 0,4452$

Portanto:

$P(50 \leq X \leq 58) = 0,4452$ ou 44,52%

Somando os dois intervalos, temos:

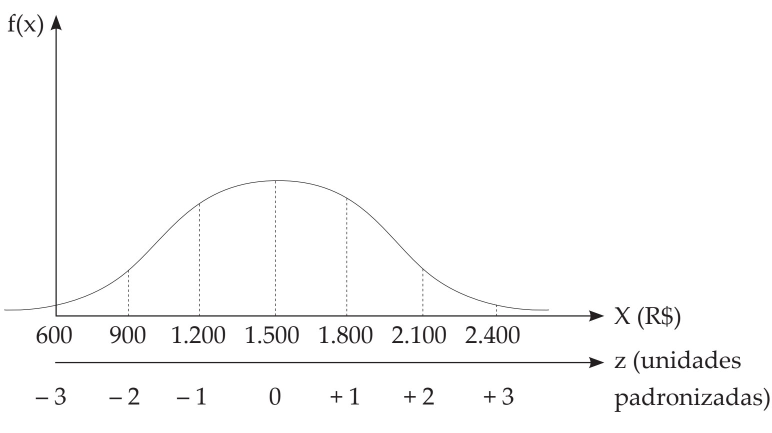
$P(48 \leq X \leq 58) = 0,1554 + 0,4452 = 0,6006$ ou 60,06% dos alunos.

Exercícios Resolvidos

1. Suponhamos que a renda média de uma grande comunidade possa ser razoavelmente aproximada por uma distribuição normal com média de R$ 1.500,00 e desvio padrão de R$ 300,00. Nessa situação:
   a) Qual a porcentagem da população que tem renda superior a R$ 1.860,00?
   b) Qual a porcentagem da população que tem renda inferior a R$ 1.000,00?
   c) Qual a porcentagem da população que tem renda entre R$ 1.050,00 e R$ 1.780,00?
   d) Numa amostra de 500 assalariados, quantos podemos esperar que tenham menos de R$ 1.050,00 de renda?

Vamos começar por fazer o gráfico (veja a Figura 7).

Figura 7 – Curva de probabilidade para o exercício 1



Vamos, então, verificar qual a porcentagem da população que tem renda superior a R$ 1.860,00.

Para X = 1.860, temos:

$$z = \frac{X - \lambda}{S} = \frac{1.860 - 1.500}{300} = +\,1{,}2$$

Então:

$P(0 \leq z \leq 1{,}2) = 0{,}3849$

Portanto:

$P(1.500 \leq X \leq 1.860) = 0{,}3849$ ou 38,49%

Nesse caso, salários maiores que R$ 1.860,00 correspondem aos valores de z maiores que 1,2. Assim:

$P(X \geq 1.860) = 0{,}5 - 0{,}3849 = 0{,}1151$ ou 11,51% da população tem renda superior a R$ 1.860,00.

Verifiquemos agora o percentual da população que tem renda inferior a R$ 1.000,00.

Para X = 1.000, temos:

$$z = \frac{X - \lambda}{S} = \frac{1.000 - 1.500}{300} = -\,1{,}67$$

Então:

$P(-1{,}67 \leq z \leq 0) = 0{,}4525$

Portanto:

$P(1.000 \leq X \leq 1.500) = 0{,}4525$ ou 45,25%

Nesse caso, salários menores que R$ 1.000,00 correspondem aos valores de z menores que – 1,67. Assim:

$P(X \leq 1.000) = 0{,}5 - 0{,}4525 = 0{,}0475$ ou 4,75% da população tem renda inferior a R$ 1.000,00.

Agora, vamos verificar qual a porcentagem da população que terá renda entre R$ 1.050,00 e R$ 1.780,00.

Para X = 1.050, temos:

$$z = \frac{X - \lambda}{S} = \frac{1.050 - 1.500}{300} = -1,5$$

Então:

P(–1,5 ≤ z ≤ 0) = 0,4332

Portanto:

P(1.050 ≤ X ≤ 1.500) = 0,4332 ou 43,32%

Para X = 1.780, temos:

$$z = \frac{X - \lambda}{S} = \frac{1.780 - 1.500}{300} = +0,93$$

Então:

P(0 ≤ z ≤ +0,93) = 0,3238

Portanto:

P(1.500 ≤ X ≤ 1.780) = 0,3238 ou 32,38%

Somando-se os dois intervalos, temos:

P(1.050 ≤ X ≤ 1.780) = 0,4332 + 0,3238 = 0,757 ou 75,7% da população.

Estamos, agora, interessados em saber, numa amostra de 500 assalariados, quantos podemos esperar que tenham renda menor que R$ 1.050,00.

Nós acabamos de calcular que para rendas entre R$ 1.050,00 e R$ 1.500,00 temos 43,32% da população. Então, abaixo de R$ 1.050,00, temos:

50% – 43,32% = 6,68%.

Resposta: Portanto, 6,68% tem renda abaixo de R$ 1.050,00.

2. Analisando os resultados das avaliações dos 2.000 alunos de certa escola, verificou-se que as notas têm uma distribuição aproximadamente normal com média igual a 6 e desvio padrão igual a 1. Quantos alunos podemos esperar que tenham tirado nota:

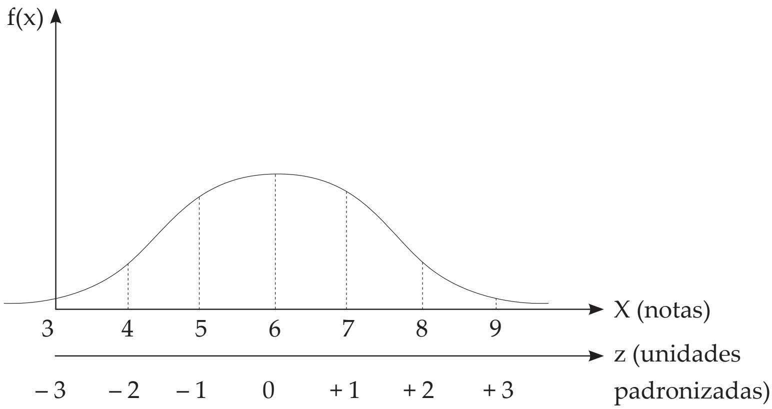
a) inferior a 5;

b) superior a 7,5;

c) entre 6,5 e 8,5.

Vamos começar pelo gráfico (veja a Figura 8).

Figura 8 – Curva de probabilidade para o exercício 2



Inicialmente, queremos saber quantos alunos tiraram nota menor que 5,0. Vamos calcular o percentual deles.

Para X = 5, temos:

$$z = \frac{X - \lambda}{S} = \frac{5 - 6}{1} = -1{,}0$$

Então:

$P(-1{,}0 \leq z \leq 0) = 0{,}3413$

Portanto:

$P(5 \leq X \leq 6) = 0{,}3413$ ou 34,13%

Como estamos interessados em notas inferiores a 5,0, queremos conhecer o percentual à esquerda de z igual a –1,0.

No caso, temos:

$0{,}5 - 0{,}3413 = 0{,}1587$ ou 15,87% dos alunos.

Como são 2.000 alunos, 15,87% deles corresponde a:

$2.000 \cdot 0{,}1587 = 317{,}4$

Resposta: Logo, esperamos encontrar 317 alunos com nota inferior a 5,0.

Não se esqueça de arredondar o resultado. Não podemos ter 0,4 aluno!

E quantos alunos devem ter tirado nota superior a 7,5?

Para X = 7,5, temos:

$$z = \frac{X - \lambda}{S} = \frac{7,5 - 6}{1} = +\,1,5$$

Então:

$P(0 \leq z \leq +1,5) = 0,4332$

Portanto:

$P(6 \leq X \leq 7,5) = 0,4332$ ou 43,32%

Mas estamos interessados nas notas superiores a 7,5. Então, estamos interessados no percentual da área sob a curva a partir de z igual a +1,5.

Logo, temos:

0,5 – 0,4332 = 0,0668 ou 6,68%

Como são 2.000 alunos, 6,68% deles corresponde a:

2.000 . 0,0668 = 133,6

Resposta: Logo, com nota superior a 7,5, esperamos encontrar 134 alunos.

Agora, vamos determinar quantos alunos tiraram nota entre 6,5 e 8,5.

Para X = 6,5, temos:

$$z = \frac{X - \lambda}{S} = \frac{6,5 - 6}{1} = +\,0,5$$

Então:

$P(0 \leq z \leq +0,5) = 0,1915$

Portanto:

$P(6 \leq X \leq 6,5) = 0,1915$ ou 19,15% dos alunos.

Para X = 8,5, temos:

$$z = \frac{X - \lambda}{S} = \frac{8,5 - 6}{1} = +\,2,5$$

Então:

$P(0 \leq z \leq +2,5) = 0,4938$

Portanto:

$P(6 \leq X \leq 8{,}5) = 0{,}4938$ ou 49,38% dos alunos.

Se 49,38% dos alunos tiraram nota entre 6,0 e 8,5 e se 19,15% dos alunos tiraram nota entre 6,0 e 6,5, então 30,23% (49,38% – 19,15% = 30,23%) dos alunos tiraram nota entre 6,5 e 8,5.

Como temos 2.000 alunos, 30,23% deles corresponde a:

$2.000 \cdot 0{,}3023 = 604{,}6$

Resposta: Logo, com nota entre 6,5 e 8,5, esperamos encontrar 605 alunos.

3. Por intermédio de documentação e observação cuidadosas, constatou-se que o tempo médio para se fazer um teste padrão de estatística é aproximadamente normal com média de 80 minutos e desvio padrão de 20 minutos. Com base nesses dados, responda:

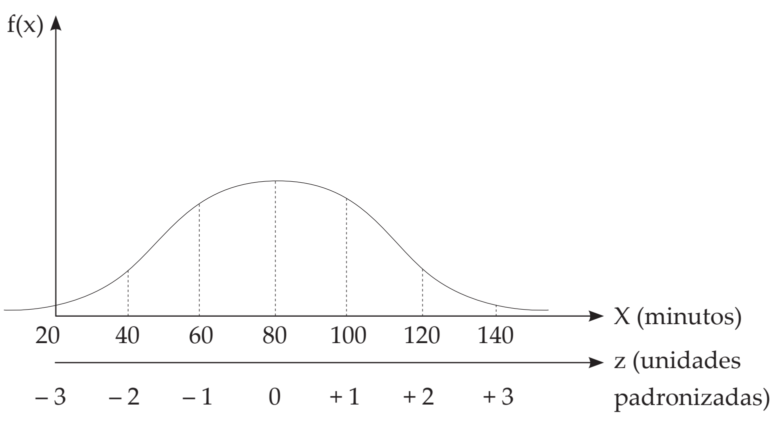
   a) Que percentual de candidatos leva menos de 80 minutos para fazer o teste?

   b) Que percentual de candidatos não termina o teste se o tempo máximo concedido for de duas horas?

   c) Se 100 pessoas fazem o teste, quantas podemos esperar que terminem o teste na primeira hora?

Vamos ao gráfico.

Figura 9 – Curva de probabilidade para o exercício 3



Vamos, então, calcular que percentual de candidatos leva menos de 80 minutos para fazer o teste.

Para X = 80, temos:

$$z = \frac{X - \lambda}{S} = \frac{80 - 80}{20} = 0$$

Resposta: Como z = 0, temos à sua esquerda 50% da curva. Logo 50% dos candidatos leva menos de 80 minutos para fazer o teste.

Agora, vamos calcular que percentual de candidatos não termina o teste se o tempo máximo concedido for de duas horas, ou seja, 120 minutos.

Para X = 120, temos:

$$z = \frac{X - \lambda}{S} = \frac{120 - 80}{20} = +2{,}0$$

Então:

$P(0 \leq z \leq +2{,}0) = 0{,}4772$

Portanto:

$P(80 \leq X \leq 120) = 0{,}4772$ ou 47,72%

Como estamos interessados em tempos maiores que 120 minutos, queremos a área abaixo da curva e à direita de z igual a +2,0.

Então, temos:

0,5 – 0,4772 = 0,0228 ou 2,28% dos candidatos.

Resposta: 2,28% dos candidatos.

Finalmente, queremos saber quantos candidatos podemos esperar que terminem o teste na primeira hora (60 minutos), supondo-se que sejam 100 pessoas a fazer o teste.

Para X = 60, temos:

$$z = \frac{X - \lambda}{S} = \frac{60 - 80}{20} = -1{,}0$$

Como queremos saber quantos candidatos levam no máximo uma hora (60 minutos) para fazer o teste, estamos interessados na área abaixo da curva e à esquerda de z igual a –1,0.

Então:

$P(-1,0 \leq z \leq 0) = 0,3413$

Portanto:

$P(60 \leq X \leq 80) = 0,3413$ ou 34,13%

Como estamos interessados em tempos menores que 60 minutos, temos:

0,5 – 0,3413 = 0,1587 ou 15,87% dos candidatos.

São 100 pessoas. Então:

100 . 0,1587 = 15,87

Resposta: Logo, 16 pessoas devem terminar o teste na primeira hora.

4. Em um exame de estatística, a média foi igual a 78 e o desvio padrão foi igual a 10. Nesse caso:

   a) determine os escores padronizados ou a variável normal padronizada de dois estudantes cujas notas foram 93 e 52;

   b) determine as notas cujos escores reduzidos foram – 0,6 e 1,2.

Para X = 93, a variável normal padronizada (z) vale:

$$z = \frac{X - \lambda}{S} = \frac{93 - 78}{10} = +1,5$$

O escore padronizado é + 1,5.

Para X = 52, a variável normal padronizada (z) vale:

$$z = \frac{X - \lambda}{S} = \frac{52 - 78}{10} = -2,6$$

O escore padronizado é – 2,6

Para z = – 0,6, o correspondente grau é:

$$z = \frac{X - \lambda}{S}$$

$$-0,6 = \frac{X - 78}{10}$$

$-6 = X - 78$

$X = 78 - 6$

Resposta: $X = 72$

Para $z = +1,2$, o grau correspondente é:

$$z = \frac{X - \lambda}{S}$$

$$1,2 = \frac{X - 78}{10}$$

$12 = X - 78$

$X = 12 + 78$

Resposta: $X = 90$

5. O processo de empacotamento em uma companhia de cereais foi ajustado de maneira que uma média de $\lambda = 13,00$ kg de cereal é colocada em cada saco. É claro que nem todos os sacos têm precisamente 13,00 kg devido a fontes aleatórias de variabilidade. O desvio padrão do peso líquido é $S = 0,1$ kg e sabemos que a distribuição dos pesos segue uma distribuição normal. Determine a probabilidade de que um saco, escolhido aleatoriamente, contenha entre 13,00 e 13,20 kg de cereal.

Para $X = 13,00$, temos:

$$z = \frac{X - \lambda}{S} = \frac{13 - 13}{0,1} = 0$$

Para $X = 13,20$, temos:

$$z = \frac{X - \lambda}{S} = \frac{13,20 - 13}{0,1} = +2,0$$

Então:

$P(0 \le z \le +2,0) = 0,4772$

Portanto:

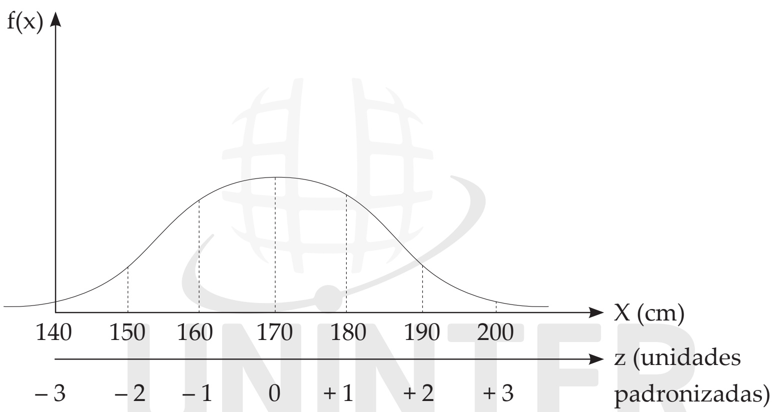
$P(13 \leq X \leq 13{,}20) = 0{,}4772$ ou 47,72%

Resposta: Ou seja, a probabilidade de que um saco escolhido aleatoriamente contenha entre 13,00 e 13,20 kg de cereal é de 47,72%.

6. As alturas dos alunos de determinada escola têm uma distribuição normal com média de 170 centímetros e desvio padrão de 10 centímetros. Qual a porcentagem de alunos dessa escola com altura entre 150 centímetros e 190 centímetros?

Vamos representar o gráfico para visualizar o problema.

Figura 10 – Curva de probabilidade para exercício 6



Para X = 150, temos:

$$z = \frac{X - \lambda}{S} = \frac{150 - 170}{10} = -2{,}0$$

Então:

$P(-2{,}0 \leq z \leq 0) = 0{,}4772$

Portanto:

$P(150 \leq X \leq 170) = 0{,}4772$ ou 47,72%

Para X = 190, temos:

$$z = \frac{X - \lambda}{S} = \frac{190 - 170}{10} = +2{,}0$$

Então:

$P(0 \leq z \leq +2{,}0) = 0{,}4772$

Portanto:

$P(170 \leq X \leq 190) = 0{,}4772$ ou 47,72%

Como estamos interessados no intervalo de 150 a 190 centímetros, precisamos somar as duas partes (de 150 a 170 e de 170 a 190). Temos, então:

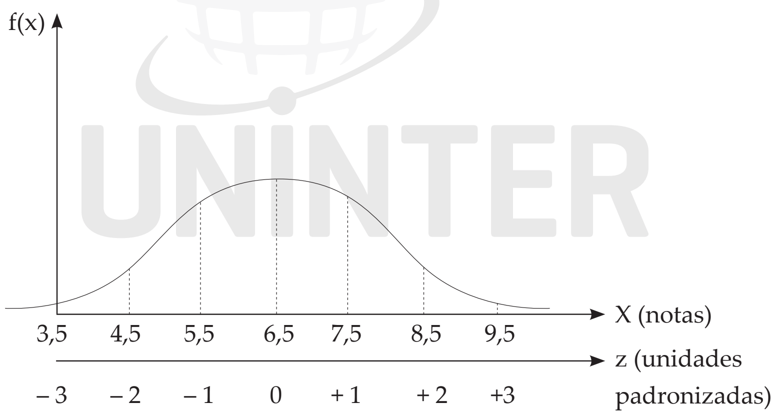
$0{,}4772 + 0{,}4772 = 0{,}9544$

Resposta: 95,44% dos alunos tem altura entre 150 e 190 centímetros.

7. Em um vestibular, verificou-se que os resultados tiveram uma distribuição normal com média igual a 6,5 e desvio padrão igual a 1,0. Qual a porcentagem de candidatos que teve média entre 4,0 e 5,0?

Vamos ao gráfico (veja a Figura 11).

Figura 11 – Curva de probabilidade para o exemplo número 10



Para X = 4,0, temos:

$$z = \frac{X - \lambda}{S} = \frac{4{,}0 - 6{,}5}{1} = -2{,}5$$

Então:

$P(-2{,}5 \leq z \leq 0) = 0{,}4938$

$$z = \frac{X - \lambda}{S} = \frac{5{,}0 - 6{,}5}{1} = -1{,}5$$

Portanto:

$P(4{,}0 \leq X \leq 6{,}5) = 0{,}4938$ ou 49,38%

Para X = 5,0, temos:

Então:

$P(-1{,}5 \leq z \leq 0) = 0{,}4332$

Portanto:

$P(5{,}0 \leq X \leq 6{,}5) = 0{,}4332$ ou 43,32%

Como estamos interessados no intervalo de notas entre 4,0 e 5,0, devemos subtrair as porcentagens encontradas, pois temos 49,38% das notas entre 4,0 e 6,5 e 43,32% das notas entre 5,0 e 6,5. Logo, entre 4,0 e 5,0, temos (49,38% – 43,32%) = 6,06%.

Resposta: Então, 6,06% dos candidatos tirou notas entre 4,0 e 5,0.

Síntese

Diferentemente de uma variável aleatória discreta, uma variável aleatória contínua pode assumir qualquer valor real, inteiro ou fracionário, dentro de um intervalo definido de valores. Dessa maneira, para distribuições de probabilidade, não conseguimos enumerar todos os possíveis valores de uma variável aleatória contínua com os valores de probabilidades correspondentes. Em lugar disso, a abordagem mais conveniente é construir uma função densidade de probabilidade ou curva de probabilidade baseada na função matemática correspondente, como foi construído neste capítulo nos vários exercícios propostos. Isso pelo fato de que a proporção da área incluída entre dois pontos quaisquer, debaixo da curva de probabilidade, identifica a probabilidade de que a variável aleatória contínua selecionada assuma um valor entre tais pontos. A área total limitada pelo eixo horizontal e pela curva normal é igual a 1 (que corresponde a 100%).

Questões para revisão

Nos exercícios a seguir, assinale a alternativa correta.

1. Em um teste de estatística realizado por 45 alunos, a média obtida foi de 5,0, com desvio padrão igual a 1,25. Determine quantos alunos obtiveram notas entre 5,0 e 7,0.

   ( ) 24 alunos
   ( ) 20 alunos
   ( ) 18 alunos
   ( ) 16 alunos
   ( ) 25 alunos

2. Uma fábrica de pneumáticos verificou que o desgaste dos seus pneus obedecia a uma distribuição normal, com média de 72.000 km e desvio padrão de 3.000 km. Calcule a probabilidade de um pneu, aleatoriamente escolhido, durar entre 69.000 km e 75.000 km.

   ( ) 34,13%
   ( ) 86,64%
   ( ) 68,26%
   ( ) 47,72%
   ( ) 43,32%

3. Uma siderúrgica verificou que os eixos de aço que fabricava para exportação tinham seus diâmetros obedecendo a uma distribuição normal, com média de 2 polegadas e desvio padrão de 0,1 polegada. Calcular a probabilidade de um eixo, aleatoriamente escolhido, ter o diâmetro com mais de 2,1 polegadas.

   ( ) 34,13%
   ( ) 15,87%
   ( ) 68,26%
   ( ) 63,48%
   ( ) 31,74%

4. As idades de um grupo de alunos apresentou média igual a vinte anos e desvio padrão igual a dois anos. Determine o percentual de alunos desse grupo que têm idade entre dezessete e vinte e dois anos.

( ) 77,45% ( ) 34,13%

( ) 43,32% ( ) 68,26%

( ) 86,64%

5. Em um vestibular, verificou-se que os resultados tiveram uma distribuição normal com média igual a 5,5 e desvio padrão igual a 1,0. Qual a porcentagem de candidatos que teve média entre 3,0 e 7,0?

( ) 49,38% ( ) 98,76%

( ) 43,32% ( ) 92,70%

( ) 86,64%

6. Uma fábrica de lâmpadas de automóveis, para exportação, verificou que a vida útil das suas lâmpadas obedecia a uma distribuição normal, com média de 2.000 horas e desvio padrão de 150 horas. Calcule a probabilidade de uma lâmpada, escolhida aleatoriamente, durar mais de 2.300 horas.

( ) 95,44% ( ) 15,87%

( ) 47,72% ( ) 2,28%

( ) 34,13%

7. A altura média dos empregados de uma empresa de seguros aproxima-se de uma distribuição normal, com média de 172 cm e desvio padrão de 8 cm. Calcule a probabilidade de um empregado dessa empresa, escolhido aleatoriamente, ter altura maior que 176 cm.

( ) 19,15% ( ) 15,87%

( ) 30,85% ( ) 38,30%

( ) 34,13%

8. Se uma amostra de 3.000 unidades de certo produto possui distribuição normal com média igual a 30, qual o desvio padrão dessa distribuição?

   DICA: Olhe, na página 172, Parâmetros da Distribuição Normal.

   ( ) 5,45 ( ) 0,99

   ( ) 29,7 ( ) 882,09

   ( ) 0,01

9. Os salários de uma empresa de *factoring* têm uma distribuição normal com média de R$ 1.800,00 e desvio padrão de R$ 180,00. Qual a probabilidade de um funcionário dessa empresa, escolhido aleatoriamente, ganhar menos de R$ 2.070,00?

   ( ) 6,68% ( ) 56,68%

   ( ) 93,32% ( ) 49,38%

   ( ) 43,32%

10. Suponha que o diâmetro médio dos parafusos produzidos por uma indústria é de 0,10 polegada, com desvio padrão de 0,01 polegada. Um parafuso é considerado defeituoso se seu diâmetro for maior que 0,11 polegada ou menor que 0,09 polegada. Qual a porcentagem de parafusos defeituosos?

    ( ) 15,87% ( ) 31,74%

    ( ) 34,13% ( ) 65,87%

    ( ) 68,26%

11. Qual a característica de uma variável aleatória contínua?

12. Defina o que é uma distribuição normal de probabilidade.

capítulo 11

A distribuição qui-quadrado

Conteúdos do capítulo

- Conceituação de distribuição qui-quadrado
- Cálculos para a aplicação da distribuição qui-quadrado

Após o estudo deste capítulo, você será capaz de:

1. definir e caracterizar a distribuição qui-quadrado;
2. aplicar a distribuição qui-quadrado.

A distribuição qui-quadrado, que representaremos por χ^2, foi estudada por Karl Pearson. Observar que χ* é letra grega *Chi* (lê-se "qui").

χ^2 é uma distribuição contínua e muito utilizada em inferência estatística.

Se temos n variáveis aleatórias independentes, representadas por $x_1, x_2, x_3, ..., x_n$, com média igual a 0 (zero) e variância igual a 1 (um), então podemos definir uma variável aleatória com distribuição qui-quadrado à variável.

$$\chi_n^2 = x_1^2 + x_2^2 + x_3^2 + ... + x_n^2$$

em que n é um parâmetro da função densidade denominado de *liberdade*. Simplificadamente,

$$\chi_n^2 = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} z_i^2 \text{, em que } x_i = \frac{X_i - \overline{X}}{S}$$

Observar que x_i é uma variável aleatória de distribuição normal reduzida.

O grau de liberdade é normalmente representado pela letra grega φ** (fi), sendo φ maior ou igual a 1.

O que significa grau de liberdade?

Cada uma das variáveis aleatórias normais atua como um número que podemos escolher livremente; e como temos n desses números, é como se tivéssemos n diferentes escolhas livres[1].

---

* A letra grega χ (qui) equivale ao ch.

** A letra grega φ (fi) equivale ao f.

Suponha uma amostra constituída por n elementos, da qual desejamos conhecer a variância (S^2).

> IMPORTANTE RECORDAR:
> S^2 representa a variância de uma amostra retirada de uma população com variância σ^2.

O primeiro passo é a determinação da média $\overline{X}$ dos n elementos. O segundo passo é a determinação de S^2.

> IMPORTANTE SABER:
> Para a população toda, a média é representada por μ e a variância por σ^2.

A função qui-quadrado de densidade é

$$f(x) = \frac{1}{c} \cdot x^{\frac{n}{2}-1} \cdot e^{-x/2}$$

Em que c é uma constante que assume valores tal que a área total sob a curva é igual a 1(um) ou 100%.

Como χ_n^2 é a soma de n variáveis aleatórias e como essas variáveis são independentes, temos que a variância é igual à soma de todas as variâncias individuais.

Então:

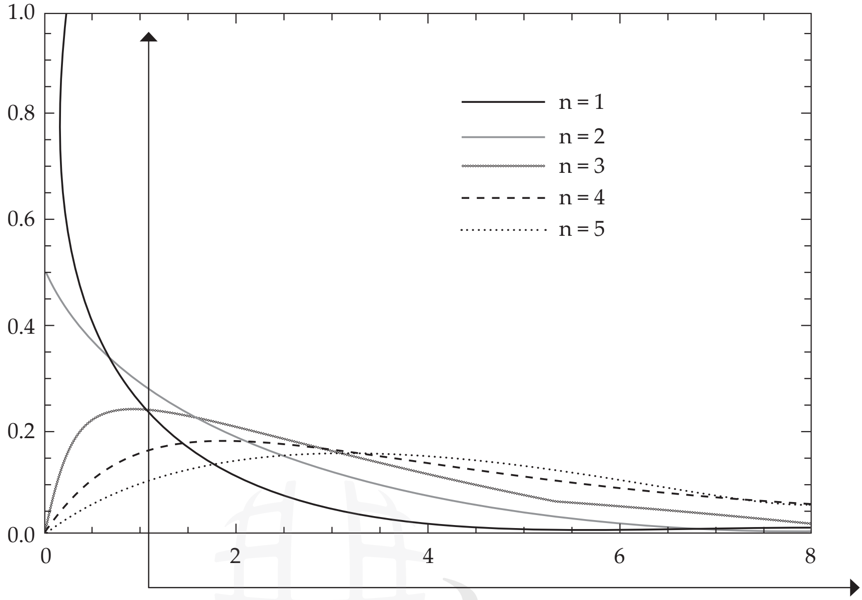
$\overline{X}_n = n$, ou seja, a média de uma distribuição qui-quadrado é igual ao grau de liberdade.

$S^2 = 2.n$, ou seja, a variância é igual ao dobro do grau de liberdade.

Para diferentes valores de n, temos diferentes curvas que descrevem a função densidade.

Vejamos alguns exemplos na Figura 12.

Figura 12 – Distribuições qui-quadrado para diferentes graus de liberdade



A área sob cada uma dessas curvas vale 1 (um) ou 100%.

Para a determinação de um percentual de área sob a curva, devemos utilizar a tabela 36.

Tabela 36 – Distribuição χ^2

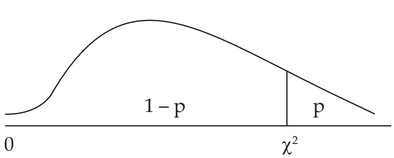
	Áreas p												
φ	0,995	0,990	0,975	0,950	0,900	0,750	0,500	0,250	0,100	0,050	0,025	0,010	0,005
1	0,0000	0,0002	0,0010	0,0039	0,0158	0,1020	0,455	1,323	2,706	3,841	5,024	6,635	7,879
2	0,0100	0,020	0,0506	0,1030	0,2110	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597
3	0,0717	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,071	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955
9	1,725	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,549	21,026	23,337	26,217	28,299
13	3,565	4,107	5,009	5,892	7,042	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819

(continua)

14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339	18,245	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338	19,369	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338	21,605	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	14,562	18,338	22,718	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337	23,828	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337	24,935	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,042	17,240	21,337	26,039	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337	27,141	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	19,037	23,337	28,241	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337	29,339	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336	30,435	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	21,749	26,336	31,528	36,741	40,113	43,194	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336	32,620	37,916	41,337	44,461	48,278	50,993
29	13,121	14,257	16,047	17,708	19,768	23,567	28,336	33,711	39,087	42,557	45,722	49,588	52,336
30	13,787	14,954	16,791	18,493	20,599	24,478	29,336	34,800	40,256	43,773	46,979	50,892	53,672
31	14,458	15,655	17,539	19,281	21,434	25,390	30,336	35,887	41,422	44,985	48,232	52,191	55,003
32	15,134	16,362	18,291	20,072	22,271	26,304	31,336	36,973	42,585	46,194	49,480	53,486	56,328
33	15,815	17,074	19,047	20,867	23,110	27,219	32,336	38,058	43,745	47,400	50,725	54,776	57,648
34	16,501	17,789	19,806	21,664	23,952	28,136	33,336	39,141	44,903	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	24,797	29,054	34,336	40,223	46,059	49,802	53,203	57,342	60,275
36	17,887	19,233	21,336	23,269	25,643	29,973	35,336	41,304	47,212	50,998	54,437	58,619	61,581
37	18,586	19,960	22,106	24,075	26,492	30,893	36,336	42,383	48,363	52,192	55,668	59,892	62,883
38	19,289	20,691	22,878	24,884	27,343	31,815	37,335	43,462	49,513	53,384	56,896	61,162	64,181
39	19,996	21,426	23,654	25,695	28,196	32,737	38,335	44,539	50,660	54,572	58,120	62,428	65,476
40	20,707	22,164	24,433	26,509	29,051	33,660	39,335	45,616	51,805	55,758	59,342	63,691	66,766
41	21,421	22,906	25,215	27,326	29,907	34,585	40,335	46,692	52,949	56,942	60,561	64,950	68,053
42	22,138	23,650	25,999	28,144	30,765	35,510	41,335	47,766	54,090	58,124	61,777	66,206	69,336
43	22,859	24,398	26,785	28,965	31,625	36,436	42,335	48,840	55,230	59,304	62,990	67,459	70,616
44	23,584	25,148	27,575	29,787	32,487	37,363	43,335	49,913	56,369	60,481	64,201	68,710	71,893
45	24,311	25,901	28,366	30,612	33,350	38,291	44,335	50,985	57,505	61,656	65,410	69,957	73,166
50	27,991	29,707	32,357	34,764	37,689	42,942	49,335	56,334	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	52,294	59,335	66,981	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	91,698	69,335	77,577	85,527	90,531	95,023	100,43	104,26
80	51,172	53,540	57,153	60,391	34,278	71,145	79,335	88,130	96,578	101,88	106,63	112,33	116,32
90	59,196	61,754	65,647	69,126	73,291	80,625	89,335	98,650	107,57	113,15	118,14	124,12	128,30
100	67,328	70,065	74,222	77,929	82,358	90,133	99,335	109,14	118,50	124,34	129,56	135,81	140,17

Para $\varphi > 30$, devemos utilizar a aproximação:

$$\chi^2 = 1/2 \cdot [\pm z_p + \sqrt{2 \cdot \varphi - 1}]^2$$

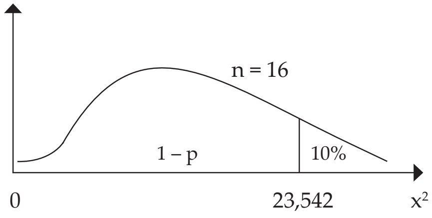


Exercício Resolvido

Desejamos calcular o valor de k tal que $p(\chi^2 > k) = 0{,}100$ com n = 16 graus de liberdade.

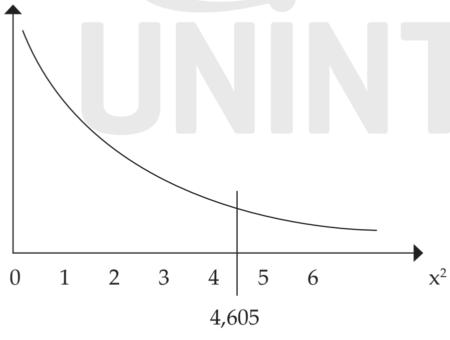
Consultando a Tabela 36, para p = 0,100 e para $\varphi = 16$, temos k = 23,542.

Graficamente, temos:



Ao traçarmos um gráfico de χ^2 em função das porcentagens acumuladas com que os valores foram encontrados, supondo grau de liberdade igual a 2, obteríamos uma distribuição como a representada na Figura 13.

Figura 13 – Distribuição qui-quadrado para grau de liberdade = 2



Vamos analisar a área delimitada pelos valores de χ^2 entre 4,605 e 9,210.

A área total sob a curva é igual a 100%. Então, o valor 4,605 delimita 10% dessa área e o valor 9,210 delimita 1% da mesma.

Então, esperamos que 90% dos resultados forneçam valores de χ^2 menores que 4,605 e que 99% dos resultados forneçam valores de χ^2 menores que 9,210.

Resposta: Logo, valores de χ^2 iguais ou maiores que 9,210 têm probabilidade de ocorrência igual a 1%.

Importante saber:
Observa-se pelas curvas que, quando o grau de liberdade é pequeno, a função de densidade é bem assimétrica e ela vai se tornando simétrica na medida em que esse grau vai aumentando. Quando n é muito grande, a distribuição qui-quadrado se assemelha a uma distribuição normal.

Síntese

A distribuição qui-quadrática, como vimos, é uma distribuição contínua e muito utilizada em inferência estatística. E, se você observar no exercício realizado neste capítulo, verá que é possível analisar pelas curvas que, quando o grau de liberdade é pequeno, a função de densidade é bem assimétrica e ela vai se tornando simétrica à medida que esse grau vai aumentando. Quando n (parâmetro da função densidade, denominada de *liberdade*) é muito grande, a distribuição qui-quadrado se assemelha a uma distribuição normal. Portanto, o que distingue a distribuição qui-quadrática é o grau de liberdade.

Questões para revisão

1. Dada a distribuição qui-quadrado com n = 12 graus de liberdade, calcule a média, a variância e o desvio padrão.

2. Calcule k para que $p(\chi^2 > k) = 0,75$ com n = 18 graus de liberdade.

3. Calcule $p(9,591 < \chi^2 < 34,170)$, com n = 20 graus de liberdade.

4. Calcule k para que $p(\chi^2 > k) = 0,75$, com n = 13 graus de liberdade.

5. Calcule k para que $p(\chi^2 > k) = 0,10$ com n = 44 graus de liberdade.

6. Assinale a alternativa correta.

   a. A distribuição qui-quadrado é representada por y^2.

   b. A distribuição qui-quadrado foi estudada por Pearson.

c. A distribuição qui-quadrado é uma distribuição discreta.

d. A distribuição qui-quadrado não serve para o uso estatístico.

7. Assinale a alternativa correta.

a. Representamos por χ^2 a distribuição qui-quadrática, sendo ela uma distribuição contínua muito utilizada nas equações lineares.

b. Grau de liberdade em distribuição qui-quadrática caracteriza-se pelo fato de cada uma das variáveis aleatórias normais atuar como um número que podemos escolher livremente.

c. Na distribuição qui-quadrática, para a população toda, a média é representada por σ e a variância por S^2.

d. S^2 representa a variância de uma amostra retirada de uma população com variância σ.

capítulo 12

Inferência estatística

Conteúdos do capítulo

- Conceituação de inferência estatística
- O processo de amostragem
- O processo de estimação
- O fator "confiança" na estatística: intervalo, nível e aplicação

Após o estudo deste capítulo, você será capaz de:

1. definir inferência estatística, amostragem, estimação e intervalo de confiança;
2. aplicar o intervalo de confiança em situações concretas.

Inferência estatística

Para realizarmos a INFERÊNCIA ESTATÍSTICA, devemos trabalhar com conhecimentos que envolvem amostragem, estimação e intervalo de confiança. Nesse processo, precisamos prestar atenção na classificação e técnicas de seleção, pois encontramos levantamentos amostrais probabilísticos e não probabilísticos, e decidirmos pelos tipos de estimativas, como veremos em detalhes na sequência.

Amostragem

Para a seleção de uma amostra que seja representativa de certa população, é necessário conhecer as técnicas utilizadas para essa seleção. Para a obtenção dos dados amostrais, os levantamentos podem ser totalmente controlados pelo pesquisador.

Tais levantamentos podem ser assim classificados:

- Levantamentos amostrais
  - Probabilísticos
    - Amostragem aleatória
      - Simples
      - Sistemática
      - Estratificada
      - Por conglomerados
  - Não probabilísticos
    - Amostragem não aleatória
      - Intencional
      - Voluntária
      - Acidental

Amostragem aleatória simples

É a forma mais fácil de se selecionar uma amostra probabilística. Todos os elementos da população têm igual probabilidade de serem selecionados.

Normalmente, atribuímos a cada elemento um número e selecionamos os números até que tenhamos os n elementos desejados da amostra.

Amostragem aleatória sistemática

É uma variação da amostragem aleatória simples e muito utilizada em pesquisas de opinião. Por exemplo, serão selecionados como amostra os alunos de uma universidade cujo nome inicie com as letras de C até J. O intervalo de amostragem de n elementos é obtido por N/n, em que N é o tamanho da população.

Amostragem aleatória estratificada

Imaginemos uma população heterogênea, constituída por todos os funcionários de uma grande indústria. Vamos dividir esses funcionários em grupos homogêneos, em função do salário dos mesmos. Cada um desses grupos homogêneos denominamos de *estrato*.

Após a identificação dos estratos, fazemos a amostragem aleatória simples de cada estrato.

Amostragem aleatória por conglomerados

Desejamos saber a escolaridade dos moradores de um bairro de determinada metrópole brasileira. Dividimos, em um mapa, esse bairro em pequenas áreas. Fazemos uma amostragem aleatória simples dessas pequenas áreas e, nas mesmas, entrevistamos todos os seus moradores para conhecermos suas escolaridades.

Amostragem não aleatória intencional

A amostra é selecionada intencionalmente pelo pesquisador. Por exemplo, o pesquisador entrevista os usuários de uma biblioteca quanto ao seu estilo preferido de leitura.

Amostragem não aleatória voluntária

Nesse caso, os elementos da população se oferecem voluntariamente para fazer parte da amostra sem a interferência do pesquisador.

Amostragem não aleatória acidental

Nesse caso, os elementos da população são escolhidos na medida em que aparecem, ou seja, são acidentalmente selecionados.

Estimação

Um dos grandes problemas da estatística é a ESTIMATIVA DAS PROPRIEDADES DAS POPULAÇÕES.

Nos capítulos 8, 9 e 10 estudamos as distribuições de probabilidades binomial, de Poisson e normal, respectivamente. Nos exemplos que analisamos, conhecíamos todas as probabilidades envolvidas e consideramos as propriedades daquelas distribuições.

Na distribuição binomial, por exemplo, a fórmula, apresentada no capítulo 8, permite-nos calcular a probabilidade de um evento acontecer se conhecemos a probabilidade p de sucesso e o valor N de tentativas. Os valores que forem atribuídos a esses dois parâmetros determinam a distribuição binomial em questão. Como normalmente determinamos o valor de N, nossos problemas de estimativas para as distribuições binomiais reduzem-se a estimar p. Na distribuição de Poisson, o problema é semelhante.

Na distribuição normal, podemos realizar qualquer cálculo se conhecermos a média (λ) e o desvio padrão (S). Então, nossos problemas de estimativas para as distribuições normais se reduzem a estimar λ e S.

Imaginemos uma amostra de 1.000 pessoas, todas eleitores. A cada pessoa dessa amostra perguntaremos se está satisfeita com o governo atual. A resposta à pergunta poderá ser SIM (satisfeita) ou NÃO (insatisfeita). Desejamos, nesse caso, estimar a proporção de pessoas que estão satisfeitas com o governo atual.

Suponhamos que 700 pessoas dessa amostra respondam SIM. Isso nos dá a estimativa natural para essa proporção, que é igual a 700/1.000, ou seja, 0,70 ou 70%. Evidentemente, estamos supondo que essa amostra representa perfeitamente a população em questão. A essa grandeza denominamos de ESTIMADOR.

> Estimador é uma grandeza baseada em observações feitas em uma amostra e que é considerada como indicador de um parâmetro populacional desconhecido.

A inferência estatística preocupa-se em conhecer as propriedades desses estimadores.

> Uma estimativa é o valor atribuído ao estimador.

No exemplo anterior, uma estimativa de p (representaremos por *E(p))* é:

E(p) = 0,7 ou 70%.

Estimativa por ponto e por intervalo

A estimativa de parâmetros na estatística pode ser por ponto ou por intervalo.

A ESTIMATIVA POR PONTO é um valor obtido a partir de cálculos efetuados com os dados da amostra, que serve como uma aproximação do parâmetro estimado.

Assim, um estimador que apresenta um único valor, como no exemplo anterior, é um estimador pontual.

Da mesma forma, se temos n elementos em uma amostra, a média amostral

$\overline{X} = (X_1 + X_2 + ... + X_n)/n$

é usada como estimador da média populacional desconhecida.

Por exemplo, se a amostra observada é constituída pelos números 3, 4, 6, 6, 8 e 9, então $\overline{X} = 6$ é uma estimativa da média populacional. Esse é, também, um estimador pontual.

A ESTIMATIVA POR INTERVALO para um parâmetro é uma faixa de valores possíveis e aceitos como verdadeiros, dentro da qual se estima que se encontre o parâmetro. A estimativa por intervalo apresenta uma maior vantagem em relação à estimativa por ponto, pois ela nos permite diminuir a magnitude do erro que estamos cometendo. Quanto menor o comprimento do intervalo, maior a precisão dos nossos cálculos.

As estimativas por intervalo são denominadas de INTERVALOS DE CONFIANÇA. Esses intervalos são baseados na distribuição amostral do estimador pontual.

Intervalo de confiança

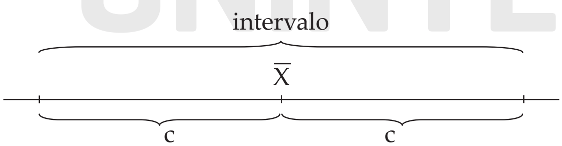
Um intervalo de confiança é um intervalo de valores obtidos a partir de observações de uma amostra e determinado de tal maneira que haja uma probabilidade de esse intervalo conter o valor desconhecido de um parâmetro que desejamos determinar. Em geral, calculamos intervalos de confiança que tenham uma chance de 95% de conter o verdadeiro valor.[1]

Nível de confiança

O nível de confiança é um número que exprime o grau de confiança (ou porcentagem) associado a um intervalo de confiança.

> IMPORTANTE: a partir de agora, chamaremos de $\overline{X}$ a média de uma amostra de tamanho n e chamaremos de μ (mu) a média de uma população qualquer. Lembre que μ não é uma variável aleatória, e, sim, um parâmetro.
>
> A partir de agora, também, chamaremos de σ* o desvio padrão de uma população e de *S* o desvio padrão da amostra.

Normalmente, $\overline{X}$ está no meio do intervalo de confiança. Assim, podemos dizer que o verdadeiro valor de μ está próximo ao valor de $\overline{X}$. Só não sabemos quão próximo. Se considerarmos que o intervalo que contém μ tem comprimento igual a 2 . c, precisamos determinar a probabilidade de μ estar entre ($\overline{X}$ – c) e ($\overline{X}$ + c). Verifique, na representação a seguir, o comprimento do intervalo de confiança:



Para um valor muito grande de c, temos certeza de que μ se encontra nesse intervalo. No entanto, um intervalo muito grande não tem aplicação prática, portanto, quanto menor o valor de c, mais preciso será o valor de μ. Há, entretanto, no caso de c pequeno, a possibilidade de que μ não se encontre no intervalo entre ($\overline{X}$ – c) e ($\overline{X}$ + c). Como já dito anteriormente, é comum escolhermos a probabilidade igual a 95% de μ estar nesse intervalo.

Então, nossos cálculos devem determinar a amplitude do intervalo para que haja 95% de chance de ele conter μ. Trata-se de um intervalo de confiança com nível de confiança igual a 95%.

Queremos, portanto, construir um intervalo de confiança (IC) da forma $\overline{X} \pm c$.

Segundo Fonseca e Martins[2], "c" é o erro amostral, expresso na unidade da variável. O erro amostral é a máxima diferença que o investigador admite suportar entre μ e $\overline{X}$.

Devemos, então, determinar o valor de c que verifique a equação:

$$P\,(\overline{X} - c < \mu < \overline{X} + c) = 0{,}95$$

Em que:

$$c = z \cdot \frac{\sigma}{\sqrt{n}}$$ para populações infinitas,

sendo que z tem distribuição normal padronizada (média igual a zero e desvio padrão igual a 1, ver capítulo n° 10) e n é o tamanho da amostra ($n \geq 30$).

Logo, para determinarmos o tamanho de uma amostra quando a população é infinita, temos que:

$$n = \left\{\frac{z \cdot \sigma}{c}\right\}^2$$

em que:

z = abscissa da distribuição normal padronizada para determinado nível de confiança; por exemplo, para nível de confiança igual a 98%, z = 2,33 (ver tabela 35 no capítulo 10)
σ = desvio padrão da população
c = erro de estimativa ou erro de amostragem; é a máxima diferença que o pesquisador admite entre $\overline{X}$ e μ
n = tamanho da amostra aleatória simples

Observe que consideramos conhecer o valor de σ. Caso não conheçamos esse valor, determinamos a média ($\overline{X}$) e o desvio padrão (S) de uma amostra aleatória e fazemos esses mesmos cálculos utilizando S no lugar de σ.[3]

No nosso exemplo, em que estamos supondo 95% de nível de confiança (em outras palavras, 5% de probabilidade de erro de estimação), z vale (ver tabela 35 no capítulo 10):

$$\frac{95\%}{2} = \frac{0{,}95}{2} = 0{,}4750, \text{ que corresponde a } z = 1{,}96$$

Por que dividimos 95% por 2?

Lembre-se de que a Tabela 35 do capítulo 10 só mostra metade da área sob a curva normal. Por isso, fizemos a divisão por 2.

Para populações finitas, temos que:

$$n = \frac{z^2 \cdot \sigma^2 \cdot N}{c^2\,(N-1) + z^2 \cdot \sigma^2}$$

Em que N é o tamanho da população.

Temos na TEORIA DA AMOSTRAGEM um teorema que diz: se a variável X possui uma distribuição normal com média μ e com desvio padrão σ, então a média $\overline{X}$, da amostra de tamanho n, também terá uma distribuição normal com média igual a μ e com desvio padrão dado por $S = \frac{\sigma}{\sqrt{n}}$

Observar que a população é infinita ou a amostragem é com reposição. Caso a população seja finita ou a amostragem seja sem reposição, temos que:

$$S = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Exemplos de aplicação de intervalos de confiança

Vamos determinar o intervalo de confiança para as pessoas de uma localidade, as quais possuem, em uma amostra de 40 pessoas, peso médio de 60 kg com desvio padrão de 3 kg. Supor nível de confiança igual a 95%.

Queremos determinar $P\,(\overline{X} - c < \mu < \overline{X} + c) = 0{,}95$.

Já vimos anteriormente que, para 95% de confiança, $z = 1{,}96$.

Então:

$$c = z \cdot \frac{\sigma}{\sqrt{n}}$$

$$c = 1{,}96 \cdot \frac{3}{\sqrt{40}}$$

$c = 0{,}93$

Logo:

$P\ (60 - 0{,}93 < \mu < 60 + 0{,}93) = 0{,}95$

O intervalo de confiança é:

$IC\ (59{,}07 < \mu < 60{,}93) = 95\%$

Isso significa que há 95% de chance de μ estar entre 59,07 kg e 60,93 kg.

Agora, vamos resolver esse exemplo, supondo nível de confiança igual a 99%.

Queremos determinar $P\ (\overline{X} - c < \mu < \overline{X} + c) = 0{,}99$.

Vamos calcular z:

$\frac{99\%}{2} = \frac{0{,}99}{2} = 0{,}4950$, que corresponde a $z = 2{,}58$

Então:

$$c = z \cdot \frac{\sigma}{\sqrt{n}}$$

$$c = 2{,}58 \cdot \frac{3}{\sqrt{40}}$$

$c = 1{,}22$

$P\ (60 - 1{,}22 < \mu < 60 + 1{,}22) = 0{,}99$

O intervalo de confiança é:

$IC\ (58{,}78 < \mu < 61{,}22) = 99\%$

Isso significa que há 99% de chance de μ estar entre 58,78 kg e 61,22 kg.

Observe que, quando aumentamos o nível de confiança para 99%, o intervalo de confiança ficou maior.

Exercícios Resolvidos

1. Determinar o intervalo de confiança para as pessoas de uma localidade, as quais possuem altura média de 168 centímetros, com desvio padrão de 20 centímetros. Supor uma amostra de 272 pessoas e nível de confiança igual a 90%.

Queremos determinar $P(\overline{X} - c < \mu < \overline{X} + c) = 0{,}90$.

Inicialmente, vamos determinar o valor de z.

$$\frac{90\%}{2} = \frac{0{,}90}{2} = 0{,}45, \text{ que corresponde a } z = 1{,}65$$

Como:

$$c = z \cdot \frac{\sigma}{\sqrt{n}}$$

Temos que:

$$c = 1{,}65 \cdot \frac{20}{\sqrt{272}}$$

$c = 2{,}0$

$P(168 - 2 < \mu < 168 + 2) = 0{,}90$

O intervalo de confiança é:

$IC(166 < \mu < 170) = 90\%$

Isso significa que há 90% de chance de μ estar entre 166 cm e 170 cm.

2. Um exportador de papel higiênico está preocupado com a metragem do seu produto. Sabemos que tal metragem tem uma distribuição aproximadamente normal, com desvio padrão de 1 metro ($\sigma = 1$). Vamos fazer um teste em um lote de 100 rolos de papel higiênico, cuja

média é de 50 metros por rolo ($\overline{X} = 50$). Determine o intervalo de confiança para um nível de confiança igual a 95,44%.

Sabemos que:

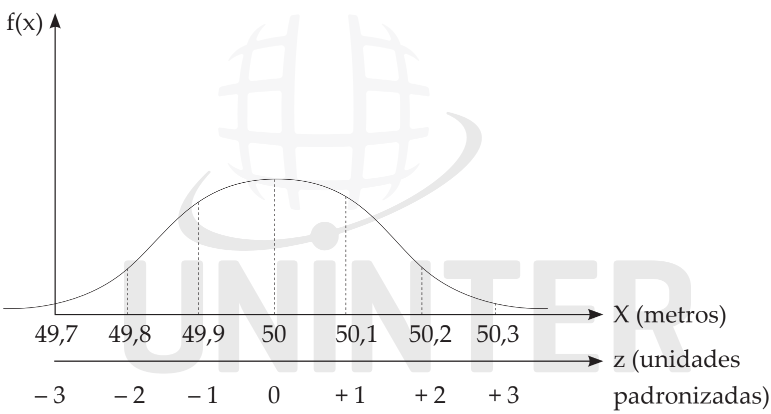
$$c = z \cdot \frac{\sigma}{\sqrt{n}}$$

Então, temos:

$$S = \frac{1}{\sqrt{100}}$$

S = 0,1 metros

Recordando a curva de distribuição normal, estudada no capítulo 10, temos que:



Como:

$$c = z \cdot \frac{\sigma}{\sqrt{n}}$$

e supondo-se um nível de confiança igual a 95,44% (z = 2, como visto anteriormente na tabela 35), há 95,44% de chance de que a variável normal assuma um valor entre $\mu - 0{,}2$ e $\mu + 0{,}2$, pois c = 2 . 0,1 ou seja c = 0,2.

No nosso exercício, $\overline{X} = 50$ (média da amostra). Então, esse valor só diverge do valor de μ (média da população) em mais de 0,2 metros para 4,56% da

produção de papel higiênico, pois em 95,44% das amostras o exportador encontra uma divergência máxima no valor de μ igual a 0,2 metros.

Resposta: o intervalo de confiança é IC $(49{,}8 < \mu < 50{,}2) = 95{,}44\%$

> UM NOVO CONCEITO PARA VOCÊ: a diferença $\overline{X} - \mu$ é denominada erro de estimativa ou erro de amostragem. Mas como saber o erro exato se um cálculo de estimativa só é feito quando μ é desconhecido? É por isso que, para o cálculo de um erro de estimativa, devemos utilizar a teoria das probabilidades apresentada no capítulo 7.

3. Considerando-se o exercício 2, qual o tamanho da amostra que se deve ter para que, com uma probabilidade de 98%, sua estimativa não esteja errada em mais de 0,15 metros?

Para 98%, temos z = 2,33 (ver tabela 35 do capítulo 10).

Então, o valor de n (nossa amostra) deve ser tal que 2,33 desvios padrão para $\overline{X}$ seja igual a 0,15. Em outras palavras:

$$2{,}33 \cdot S = 0{,}15$$

Como $S = \dfrac{\sigma}{\sqrt{n}}$ e $\sigma = 1$

temos:

$$2{,}33 \cdot \frac{1}{\sqrt{n}} = 0{,}15$$

$$\sqrt{n} = \frac{2{,}33}{0{,}15} = 15{,}533$$

$$n = 15{,}533^2$$

Resposta: n = 241 rolos de papel higiênico

4. Suponhamos que, no exercício anterior, estivéssemos preocupados com um nível de confiança de apenas 70%. Nesse caso, qual é o tamanho da amostra?

70% corresponde a um z = 1,04.

Temos, então:

$$1{,}04 \cdot \frac{1}{\sqrt{n}} = 0{,}15$$

$$n = \left(\frac{1{,}04}{0{,}15}\right)^2$$

Resposta: n = 48 rolos de papel higiênico.

5. Considerando ainda o mesmo exercício, qual o intervalo de confiança para nível de confiança igual a 97% e uma amostra de 100 rolos de papel?

Para nível de confiança igual a 97%, o valor de z é 2,17 (ver tabela 35).

Como a população é infinita, $c = z \cdot \frac{\sigma}{\sqrt{n}}$

$$c = 2{,}17 \cdot \frac{1}{\sqrt{100}}$$

c = 0,217

$P\ (50 - 0{,}217 < \mu < 50 + 0{,}217) = 0{,}97$

O intervalo de confiança é:

$IC\ (49{,}783 < \mu < 50{,}217) = 97\%$

Resposta: Isso significa que há 97% de chance de μ estar entre 49,783 m e 50,217 m.

6. Os dados a seguir relacionam-se ao consumo de energia de cem residências em um bairro de baixa renda de Curitiba, em kWh (quilowatts hora).

45 55 48 56 70 66 88 58 46 55

76 41 67 84 40 59 67 82 53 90

53 86 74 75 43 58 91 83 76 50

44 70 85 49 51 43 76 77 85 57

67 69 75 46 47 58 60 64 80 49

61 66 90 87 45 46 88 74 56 50

44 75 78 93 45 82 81 41 55 58

60 60 53 78 64 49 63 71 70 44

77 54 69 72 90 73 57 50 55 60

51 88 85 45 66 78 46 54 57 59

Já calculamos para você a média e o desvio padrão dessa população, que são os valores:

$\mu = 64$

$\sigma = 14,82$

Agora retire uma amostra contendo três linhas (30 elementos) quaisquer e monte uma distribuição de frequências. A seguir, calcule a média e o desvio padrão dessa amostra. Finalmente, determine o intervalo de confiança para 97% de nível de confiança.

Para que todos cheguem ao mesmo resultado, sugerimos que escolham as três primeiras linhas dos dados, ou seja:

45 55 48 56 70 66 88 58 46 55

76 41 67 84 40 59 67 82 53 90

53 86 74 75 43 58 91 83 76 50

Vamos, então, MONTAR A DISTRIBUIÇÃO DE FREQUÊNCIAS (uma tabela).

Dados	f
40	1
41	1
43	1
45	1
46	1
48	1
50	1
53	2
55	2
56	1
58	2
59	1
66	1
67	2
70	1
74	1
75	1
76	2
82	1
83	1
84	1
86	1
88	1
90	1
91	1
Σ 1935	30

A MÉDIA DA AMOSTRA É:

$$\overline{X} = \frac{1935}{30} = 64{,}5$$

Para nível de confiança de 97% temos:

$$\frac{97\%}{2} = \frac{0{,}97}{2} = 0{,}485 \text{ que corresponde a } z = 2{,}17$$

Como:

$$c = z \cdot \frac{\sigma}{\sqrt{n}}$$

$c = 2{,}87 \cdot \frac{14{,}82}{\sqrt{30}}$

$c = 5{,}87$

$P\,(64{,}5 - 5{,}87 < \mu < 64{,}5 + 5{,}87) = 0{,}97$

O intervalo de confiança é:

$IC\,(58{,}63 < \mu < 70{,}37) = 97\%$

Resposta: Isso significa que há 97% de chance de m estar entre 58,63 kWh e 70,37 kWh.

7. Calcule o intervalo de confiança no exercício anterior, para nível de confiança igual a 85%.

$\frac{85\%}{2} = \frac{0{,}85}{2} = 0{,}425$, que corresponde a $z = 1{,}44$

Como:

$c = z \cdot \frac{\sigma}{\sqrt{n}}$

Temos que:

$c = 1{,}44 \cdot \frac{14{,}82}{\sqrt{30}}$

$c = 3{,}9$

$P\,(64{,}5 - 3{,}9 < \mu < 64{,}5 + 3{,}9) = 0{,}85$

O intervalo de confiança é:

$IC\,(60{,}6 < \mu < 68{,}4) = 85\%$

Resposta: Isso significa que há 85% de chance de m estar entre 60,6 kWh e 68,4 kWh.

8. Justifique o fato de no exercício 6 o intervalo de confiança ser maior que o intervalo do exercício 7.

No exercício 6, exigiu-se um nível de confiança de 97% contra apenas 85% no exercício 7. Em função disso, o intervalo no 6 é maior para aumentar a chance de μ encontrar-se dentro do mesmo.

Síntese

A INFERÊNCIA ESTATÍSTICA implica admitirmos que os resultados obtidos na análise dos dados de uma amostra são válidos para toda a população da qual aquela amostra foi retirada. Consiste em obtermos e generalizarmos conclusões sobre dada característica de uma população a partir de informações colhidas da amostra. Já havíamos estudado esses conceitos no capítulo 1 e voltamos a eles para fazermos uma sequência de construção de conhecimentos que nos permitissem aplicarmos em situações de elaborações estatísticas os processos e concepções de:

- Amostragem
- Tamanho de uma amostra
- Estimação
- Estimativa por ponto e por intervalo
- Intervalo de confiança
- Nível de confiança

Procedimentos esses que concluímos com exemplos de aplicação de intervalos de confiança.

Questões para revisão

Assinale a alternativa correta nos exercícios a seguir.

1. Determine o intervalo de confiança para as pessoas de uma localidade onde os habitantes possuem peso médio de 68 kg com desvio padrão de 3 kg. Supor nível de confiança igual a 90% e uma amostra de 64 pessoas.

   ( ) IC $(67,38 < \mu < 68,62) = 90\%$

   ( ) IC $(67,92 < \mu < 68,08) = 90\%$

   ( ) IC $(63,05 < \mu < 72,95) = 90\%$

   ( ) IC $(63,60 < \mu < 72,40) = 90\%$

   ( ) IC $(66,35 < \mu < 69,65) = 90\%$

2. Determine o intervalo de confiança para as pessoas de uma localidade onde os habitantes possuem altura média de 162 centímetros, com desvio padrão de 18 centímetros. Supor uma amostra de 138 pessoas e nível de confiança igual a 95%.

   ( ) IC ($153 < \mu < 171$) = 95%

   ( ) IC ($126,72 < \mu < 197,28$) = 95%

   ( ) IC ($156 < \mu < 168$) = 95%

   ( ) IC ($144 < \mu < 180$) = 95%

   ( ) IC ($159 < \mu < 165$) = 95%

3. Considere os dados sobre o consumo de energia do exercício resolvido 6. Em seguida, considere a amostra constituída pelas três últimas linhas (30 elementos), ou seja:

   60 60 53 78 64 49 63 71 70 44

   77 54 69 72 90 73 57 50 55 60

   51 88 85 45 66 78 46 54 57 59

   Determine o intervalo de confiança para 99% de nível de confiança.

   Lembre-se que conhecemos a média e o desvio padrão dessa população. Os valores são:

   $\mu = 64$ e $\sigma = 14,82$. Calcule o valor de "c" com duas casas após a vírgula.

   ( ) IC ($56,29 < \mu < 70,25$) = 99%

   ( ) IC ($61,72 < \mu < 66,88$) = 99%

   ( ) IC ($62,39 < \mu < 66,21$) = 99%

   ( ) IC ($50,70 < \mu < 77,90$) = 99%

   ( ) IC ($57,80 < \mu < 70,80$) = 99%

4. No exercício 3, qual seria o intervalo de confiança para um nível de confiança de 75,80%?

   ( ) IC $(57{,}80 < \mu < 70{,}80) = 75{,}80\%$

   ( ) IC $(46{,}96 < \mu < 81{,}64) = 75{,}80\%$

   ( ) IC $(62{,}39 < \mu < 66{,}21) = 75{,}80\%$

   ( ) IC $(60{,}10 < \mu < 66{,}44) = 75{,}80\%$

   ( ) IC $(50{,}70 < \mu < 77{,}90) = 75{,}80\%$

5. Determine o intervalo de confiança para os empregados de uma empresa que possuem salário médio de R$ 1.840,00 com desvio padrão de R$ 300,00. Supor nível de confiança igual a 95% e uma amostra de 96 empregados.

   ( ) IC $(1540 < \mu < 2140) = 95\%$

   ( ) IC $(1252 < \mu < 2428) = 95\%$

   ( ) IC $(1780 < \mu < 1900) = 95\%$

   ( ) IC $(1600 < \mu < 2080) = 95\%$

   ( ) IC $(1744 < \mu < 1936) = 95\%$

6. De acordo com o que vimos aqui, como você define o processo de inferência estatística?

7. O que é um intervalo de confiança?

capítulo 13

Teste de hipóteses

Conteúdos do capítulo

- Conceituação de teste de hipóteses
- Os tipos de hipóteses e etapas de um teste de hipóteses
- Os níveis de significância e as regiões de rejeição e aceitação das hipóteses.

Após o estudo deste capítulo, você será capaz de:

1. definir teste de hipóteses e seus fatores;
2. aplicar teste de hipóteses.

Teste de hipóteses

O que é teste de hipóteses?

É uma técnica para se fazer inferência estatística.[1]

Isso significa dizer que, a partir de um teste realizado com os dados de uma amostra, pode-se inferir sobre a população a que essa amostra pertence.

Lembre-se que, quando desejamos avaliar um parâmetro populacional sobre o qual não temos qualquer informação do seu valor, temos que o estimar por intermédio do intervalo de confiança. Mas, se conhecemos o valor desse parâmetro que desejamos avaliar, podemos testá-lo para verificar se é verdadeiro ou falso.

Uma hipótese estatística é, portanto, uma suposição quanto ao valor de um parâmetro populacional. O teste de hipóteses é uma técnica que nos permite aceitar ou rejeitar a hipótese estatística, a partir dos dados da amostra dessa população.

Vamos iniciar nossa análise a partir de uma moeda.

Ao jogarmos uma moeda, admitimos que a mesma é honesta (ou equilibrada), isto é, em qualquer jogada há igual probabilidade de obtermos cara (50%) ou de obtermos coroa (50%).

Mas como podemos ter certeza de que a moeda é realmente honesta ou equilibrada? Suponhamos que, ao jogar uma moeda, desejamos obter o resultado "cara".

Façamos inicialmente uma constatação óbvia: se a moeda tem duas caras, então $p = 1$; se a moeda tem duas coroas, então $p = 0$. Continua difícil afirmar que a moeda é honesta ou equilibrada. Não basta olhar para a moeda e a jogarmos apenas uma vez: é impossível afirmar que ela é ou não honesta ou equilibrada. Faz-se necessário jogá-la um grande número de vezes e observar os resultados para termos alguma base de julgamento.

Hipótese nula e hipótese alternativa

Hipótese nula (H_0) é a informação (a hipótese) que será testada. É a informação a respeito do valor do parâmetro que desejamos avaliar.[2]

Hipótese alternativa (H_1) é a hipótese que afirma que a hipótese nula é falsa. É a afirmação a respeito do valor do parâmetro que aceitaremos como verdadeiro, caso a hipótese nula seja rejeitada.

Observe que a hipótese nula é uma igualdade, enquanto a hipótese alternativa é uma desigualdade.

Por exemplo:

a) H_0: $\mu = 70$ kg

   H_1: $\mu > 70$ kg;

b) H_0: $\mu = 70$ kg

   H_1: $\mu < 70$ kg;

c) H_0: $\mu = 70$ kg

   H_1: $\mu \neq 70$ kg.

Continuemos a análise da nossa moeda, para melhorarmos o entendimento de hipótese nula e de hipótese alternativa.

Primeiro, decidimos quanto à hipótese a ser testada. No caso da moeda (anteriormente comentada), a hipótese é que $p = 0,5$. A hipótese que será testada, como já dissemos, é denominada de *hipótese nula* e a hipótese que afirma que a hipótese nula é falsa é denominada de *hipótese alternativa*.

No nosso exemplo, a hipótese alternativa é que a moeda não seja honesta ou equilibrada ($p \neq 0,5$).

Como só temos duas possibilidades, sabemos que ou a hipótese nula ou a hipótese alternativa é verdadeira. Nosso problema consiste em aceitar a hipótese nula e dizer que a moeda é honesta ou equilibrada ou rejeitar a hipótese nula e afirmar que a moeda não é honesta ou equilibrada. O QUE FAZER?

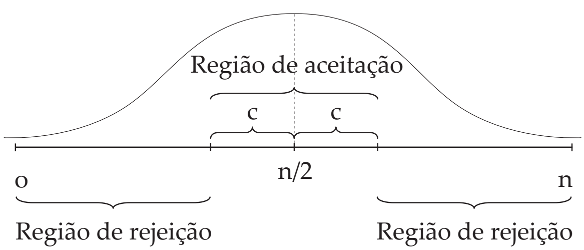
Intuitivamente sabemos que devemos jogar a moeda um grande número de vezes (vamos supor n vezes) e verificar os resultados obtidos (cara ou coroa). Caso o número de caras esteja próximo a n/2, aceitamos a hipótese de que a moeda é honesta ou equilibrada. Caso contrário (número de caras muito diferente de n/2), rejeitamos tal hipótese. Aí, vem a pergunta: quão diferente de n/2 o resultado deve ser para que possamos aceitar ou rejeitar a hipótese de que a moeda é honesta ou equilibrada?

Como dissemos anteriormente, o teste de hipóteses é uma técnica que nos permite aceitar ou rejeitar a hipótese estatística, a partir dos dados da amostra dessa população. Logo, usaremos uma amostra da população em estudo, para verificar se ela confirma ou não o valor do parâmetro informado pela hipótese nula. Entretanto, pela aceitação ou não de uma hipótese nula, podemos cometer erros na decisão.

VOLTEMOS À NOSSA MOEDA.

Nosso processo de teste é o seguinte: escolhemos um número "c" e, se o número de caras está entre (n/2 – c) e (n/2 + c), aceitamos a hipótese nula, bem como concluímos que a moeda é honesta ou equilibrada; caso contrário, dizemos que a moeda não é honesta.

A região de (n/2 – c) a (n/2 + c) é denominada de *região de aceitação* e a região para a qual a hipótese nula será rejeitada é denominada de *região de rejeição* ou de *região crítica* (veja a ilustração a seguir). Mas o nosso problema continua, pois não sabemos quão grande pode ser o valor de c.



Erros do tipo 1 e do tipo 2

Há várias formas de realizarmos o teste de hipóteses. Entretanto, corremos o risco de rejeitar uma hipótese que é verdadeira ou de aceitá-la quando é falsa. Temos, portanto, dois tipos possíveis de erros que são representados no quadro a seguir.

	Aceita-se a hipótese nula (H_o)	Rejeita-se a hipótese nula (H_o)
H_o é verdadeira.	Decisão foi correta.	Erro do tipo 1.
H_o é falsa.	Erro do tipo 2.	Decisão foi correta.

Para melhor compreendermos esse quadro, voltemos ao caso, visto anteriormente, da moeda.

Nós sempre desejamos acertar ao fazermos um julgamento sobre nossa hipótese nula. Entretanto, só há duas maneiras de acertarmos: aceitando a hipótese nula, quando ela é verdadeira, ou rejeitando-a, quando ela é falsa.

Logo há duas possibilidades de errarmos:

rejeitando a hipótese nula, quando ela for verdadeira;

ou aceitando-a, quando ela for falsa.

Ao primeiro tipo de erro, denominamos de *tipo 1*; ao segundo tipo de erro, denominamos de *tipo 2*.

Naturalmente, nós nos preocupamos mais com a possibilidade de rejeitar de forma equivocada a hipótese nula, ou seja, temos maior cuidado em evitar o erro do tipo 1.

Normalmente, fixa-se um limite superior para a probabilidade de cometermos um erro do tipo 1. O valor mais comum para esse limite é de 5%, sendo que em alguns casos, aceita-se 10%.

Nível de significância

Nível de significância do teste é o nome dado à probabilidade de cometermos um erro do tipo 1. Dizemos, por exemplo, que o nosso teste de hipóteses está planejado com nível de significância igual a 5%. Vamos chamá-lo de α (alfa)* (no caso, α = 5%).

* α é a letra grega minúscula que lemos *alfa* e que corresponde à letra *a* do nosso alfabeto.

Observe que um erro do tipo 1 é, socialmente falando, mais importante que um erro do tipo 2. Vejamos alguns exemplos práticos de análise do nível de significância.

Vamos, NESTA PRIMEIRA SITUAÇÃO, supor a decisão de aprovar ou reprovar um aluno na prova final.

	Aprovar o aluno	Reprovar o aluno
O aluno tirou nota mínima.	Decisão correta.	Erro do tipo 1.
O aluno não tirou nota mínima.	Erro do tipo 2.	Decisão correta.

Agora, UMA SEGUNDA SITUAÇÃO: suponhamos que um médico precise decidir se vai ou não retirar um tumor benigno de um paciente.

	Opera o paciente	Não opera o paciente
O paciente precisa ser operado.	Decisão correta.	Erro do tipo 1.
O paciente não precisa ser operado.	Erro do tipo 2.	Decisão correta.

Como chamamos de α (alfa) a probabilidade de cometer um erro do tipo 1, chamaremos de β (beta)* a probabilidade de cometer um erro do tipo 2.

Região de rejeição e região de aceitação

Para a análise da região de rejeição e da região de aceitação, precisamos fazer três considerações.

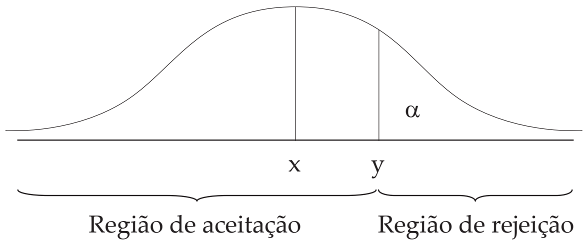
a) H_0: parâmetro = x

H_1: parâmetro > x

NESSA PRIMEIRA CONSIDERAÇÃO, se uma amostra aleatória fornecer uma estimativa menor ou igual a x, devemos aceitar a hipótese nula (H_0). Caso H_0 seja verdadeira, tomamos a decisão correta. Mas, se a amostra aleatória fornecer uma estimativa muito maior que x, provavelmente rejeitaremos H_0. Assim, se a hipótese nula for verdadeira, teremos cometido um erro do tipo 1.

Por exemplo, se y é a distribuição amostral do parâmetro, então a região de aceitação da hipótese nula é a região à esquerda de y e a região de rejeição é a região à direita de y, conforme ilustrado a seguir.

* β é a letra grega minúscula que lemos *beta* e que corresponde à letra *b* do nosso alfabeto.

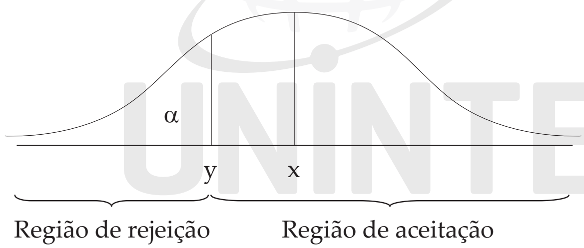


b) H_0: parâmetro = x

H_1: parâmetro < x

Nessa segunda consideração, se uma amostra aleatória fornecer uma estimativa maior ou igual a x, devemos aceitar a hipótese nula (H_0). Caso H_0 seja verdadeira, tomamos a decisão correta. Mas, se a amostra aleatória fornecer uma estimativa muito menor que x, provavelmente rejeitaremos H_0. Assim, se a hipótese nula for verdadeira, teremos cometido um erro do tipo 1.

Por exemplo, se y é a distribuição amostral do parâmetro, então a região de aceitação da hipótese nula é a região à direita de y e a região de rejeição é a região à esquerda de y, conforme ilustrado a seguir.

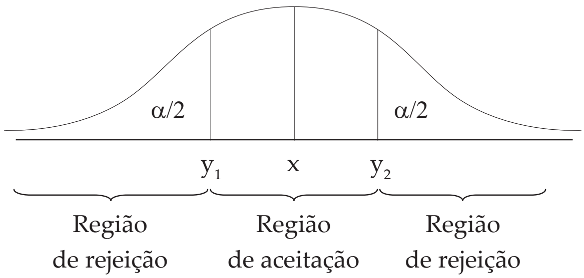


c) H_0: parâmetro = x

H_1: parâmetro ≠ x

Nessa terceira consideração, se uma amostra aleatória fornecer uma estimativa próxima a x, devemos aceitar a hipótese nula (H_0). Caso H_0 seja verdadeira, tomamos a decisão correta. Mas, se a amostra aleatória fornecer uma estimativa ou muito maior que x ou muito menor que x, provavelmente rejeitaremos H_0. Assim, se a hipótese nula for verdadeira, teremos cometido um erro do tipo 1.

Por exemplo, se y_1 ou y_2 é a distribuição amostral do parâmetro, então a região de aceitação da hipótese nula é a região à esquerda de y_2 ou à direita de y_1, conforme ilustrado a seguir.



Etapas de um teste de hipóteses

Para qualquer teste de hipóteses, devemos observar as seguintes etapas:

a) enunciar a hipótese a ser testada (H_0) e enunciar qual a hipótese alternativa (H_1);

b) a partir de H_1, definir o tipo de teste que será usado para testar H_0;

c) fixar o limite de erro (α), ou seja, fixar a probabilidade de cometer-se um erro do tipo 1;

d) determinar a região de rejeição e a região de aceitação;

e) com os elementos amostrais, calcular o estimador e verificar se ele se encontra ou na região de rejeição ou na região de aceitação;

f) decidir, se o estimador estiver na região de aceitação, aceitar H_0; se o estimador estiver na região de rejeição, rejeitar H_0.

Caso não conheçamos σ (desvio padrão populacional), usaremos S (desvio padrão amostral).

Teste para médias

Um ótimo estimador para a média populacional (μ) é a média amostral ($\overline{X}$). A distribuição amostral das médias é normal, com:

$$Z_r = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

em que Z_r é denominado teste Z de uma amostra.

Vamos analisar um exemplo de distribuição amostral de médias

Primeiramente analisaremos a situação.

$H_0: \mu = x$

$H_1: \mu > x$

Suponhamos uma amostra aleatória de 100 elementos, com média igual a 70, retirados de uma população normal com desvio padrão $\sigma = 5$. Considerando-se um nível de significância de 5%, teste a hipótese de que a média populacional (μ) seja igual a 69. Suponha a hipótese alternativa $\mu > 69$.

Vamos determinar a região de rejeição para $\alpha = 5\%$.

Consultando a tabela 35 do capítulo 10, verificamos que, para 5% temos z = 1,65, ou seja, 50% – 5% = 45% = 0,45.

Vamos, agora, determinar z para a região de rejeição (z_r).

$$Z_r = \frac{70 - 69}{\frac{5}{\sqrt{100}}}$$

$$Z_r = \frac{1}{\frac{5}{10}}$$

$$Z_r = 2$$

Verificamos que o valor de Z_R está na região de rejeição para a hipótese nula, pois $Z_R > 1,65$. Portanto, a hipótese nula será rejeitada.

Exercícios Resolvidos

1. $H_0: \mu = x$

   $H_1: \mu < x$

Suponhamos uma amostra aleatória de 81 elementos, com média igual a 92, retirados de uma população normal com desvio padrão $\sigma = 5$. Considerando um nível de significância de 10%, teste a hipótese de que a média populacional (μ) seja igual a 93. Suponha a hipótese alternativa $\mu < 93$.

Vamos determinar a região de rejeição para α = 10%.

Consultando a tabela 35 do capítulo 10, verificamos que, para 10%, temos z = – 1,28, ou seja: 50% - 10% = 40% = – 0,40.

Vamos agora determinar z para a região de rejeição (Z_r).

$$Z_r = \frac{92 - 93}{\frac{5}{\sqrt{81}}}$$

$$Z_r = \frac{-1}{\frac{5}{9}}$$

$$Z_r = -1{,}80$$

Verificamos que o valor de Z_R está na região de rejeição para a hipótese nula, uma vez que $Z_R < -1{,}28$. Portanto, a hipótese nula será rejeitada.

2. $H_0: \mu = x$

   $H_1: \mu > x$

Suponhamos uma amostra aleatória de 144 elementos, com média igual a 68, retirados de uma população normal com desvio padrão σ = 8. Nesse caso, considerando um nível de significância de 5%, teste a hipótese de que a média populacional (μ) seja igual a 67. Suponha a hipótese alternativa μ > 67.

Vamos determinar a região de rejeição para α = 5%.

Consultando a tabela 35 do capítulo 10, verificamos que, para 5%, temos z = 1,65, ou seja, 50% – 5% = 45% = 0,45.

Vamos agora determinar z para a região de rejeição (Zr).

$$Z_r = \frac{68 - 67}{\frac{8}{\sqrt{144}}}$$

$$Z_r = \frac{1}{\frac{8}{12}}$$

$$Z_r = 1{,}5$$

VERIFICAMOS QUE O VALOR DE Z_R ESTÁ NA REGIÃO DE ACEITAÇÃO PARA A HIPÓTESE NULA, POIS $Z_R < 1,65$. PORTANTO, A HIPÓTESE NULA SERÁ ACEITA.

Síntese

Uma vez que consideramos ser uma hipótese estatística uma suposição quanto ao valor de um parâmetro populacional, vimos, neste capítulo, que utilizamos o teste de hipóteses por ser uma técnica que nos possibilita aceitar ou rejeitar a hipótese estatística a partir dos dados da amostra dessa população. E, para realizarmos esse processo, precisamos de clareza quanto ao que seja:

- Hipótese nula e hipótese alternativa
- Erros do tipo 1 e do tipo 2
- Nível de significância
- Região de rejeição e região de aceitação
- Etapas de um teste de hipóteses
- Teste para médias

Assim, conceituamos todos esses aspectos, para em seguida aplicá-los em situações concretas que os habilitam para realizar uma inferência estatística: o teste de hipóteses.

Questões para revisão

Marque a alternativa correta nos exercícios a seguir.

1. Temos uma amostra aleatória de 64 elementos, com média igual a 50, retirados de uma população normal com desvio padrão $\sigma = 6$. Considerando um nível de significância de 5%, teste a hipótese de que a média populacional (μ) seja igual a 52, sendo a hipótese alternativa $\mu < 52$.

   ( ) $z_r = 1{,}65$ e está na região de aceitação.

   ( ) $z_r = -1{,}65$ e está na região de aceitação.

   ( ) $z_r = 2{,}67$ e está na região de aceitação.

   ( ) $z_r = -2{,}67$ e está na região de rejeição.

   ( ) $z_r = -1{,}65$ e está na região de rejeição.

2. Temos uma amostra aleatória de 100 elementos, com média igual a 88, retirados de uma população normal com desvio padrão $\sigma = 20$. Considerando um nível de significância de 5%, teste a hipótese de que a média populacional (μ) seja igual a 85, sendo a hipótese alternativa $\mu > 85$.

   ( ) $z_r = -1{,}65$ e está na região de rejeição.

   ( ) $z_r = -1{,}5$ e está na região de rejeição.

   ( ) $z_r = 1{,}5$ e está na região de aceitação.

   ( ) $z_r = -1{,}5$ e está na região de aceitação.

   ( ) $z_r = 1{,}5$ e está na região de rejeição.

3. Temos um empacotador automático de café que funciona de maneira que a quantidade de café em cada pacote de 500 gramas tenha uma distribuição normal com variância igual a 25. Considerando um nível de significância de 5%, teste a hipótese de que a média μ seja igual a 500, sendo a hipótese alternativa $\mu > 500$. Foram dadas dez amostras com os seguintes pesos: 508, 510, 494, 500, 505, 511, 508, 499, 496, 489.

( ) A hipótese que $\mu = 500$ é aceita, pois $z_r < 1,65$.

( ) A hipótese que $\mu = 500$ é rejeitada, pois $z_r < 1,65$.

( ) A hipótese que $\mu = 500$ é aceita, pois $z_r > 1,65$.

( ) A hipótese que $\mu = 500$ é rejeitada, pois $z_r > 1,65$.

( ) A hipótese que $\mu = 500$ é rejeitada, pois $z_r = 1,65$.

4. Temos uma amostra aleatória de 40 elementos, com média igual a 100, retirados de uma população normal com desvio padrão $\sigma = 12$. Considerando um nível de significância de 10%, teste a hipótese de que a média populacional (μ) seja igual a 102, sendo a hipótese alternativa $\mu < 102$.

   ( ) A hipótese nula será rejeitada, porque Z_r está na região de aceitação.

   ( ) A hipótese nula será rejeitada, porque Z_r está na região de rejeição.

   ( ) A hipótese nula será aceita, porque Z_r está na região de aceitação.

   ( ) A hipótese nula será aceita, porque Z_r está na região de rejeição.

   ( ) A hipótese nula será rejeitada, porque não se pode determinar Z_r.

5. Temos uma amostra aleatória de 30 elementos, com média igual a 48, retirados de uma população normal com desvio padrão $\sigma = 10$. Considerando um nível de significância de 10%, teste a hipótese de que a média populacional (μ) seja igual a 46, sendo a hipótese alternativa $\mu > 46$.

   ( ) A hipótese nula será rejeitada, porque Z_r está na região de aceitação.

   ( ) A hipótese nula será rejeitada, porque Z_r está na região de rejeição.

   ( ) A hipótese nula será aceita, porque Z_r está na região de aceitação.

   ( ) A hipótese nula será aceita, porque Z_r está na região de rejeição.

   ( ) A hipótese nula será rejeitada, porque não se pode determinar Z_r.

6. O que é uma hipótese nula e uma hipótese alternativa?

7. O que significa no teste de hipóteses "nível de significância"?

capítulo 14

Análise da variância (Anova)

Conteúdos do capítulo

- Conceituação do método de cálculo da variância – Anova
- Aplicação da Anova em situações práticas

Após o estudo deste capítulo, você será capaz de:

1. conceituar o método de variância – Anova;
2. aplicar o cálculo de variância em situações práticas.

Já estudamos no teste de hipóteses como verificar a igualdade entre duas médias. Entretanto, se precisarmos comparar três ou mais médias, simultaneamente, a comparação das mesmas, duas a duas, seria um trabalho pouco eficiente.

Para solucionar esse problema, Fisher* criou a denominada *Análise da Variância* (abreviadamente, *Anova = Analysis of Variance*), que nos permite comparar, simultaneamente, as médias de várias amostras, desde que:

a) tais amostras tenham sido extraídas de populações que têm distribuição normal;

b) as populações tenham o mesmo valor de variância;

c) tais amostras sejam aleatórias e independentes.

O método de análise da variância consiste em dividir a variância em componentes úteis. O modelo mais simples de Anova é aquele em que observaremos os grupos considerando uma única propriedade.

Suponhamos que desejamos analisar as médias obtidas em Estatística pelos alunos de três diferentes localidades que assistem às aulas na modalidade a distância.

Vamos, então, considerar os seguintes resultados amostrais:

Localidade A: 8 – 7 – 8 – 6 – 9 – 7 – 6 – 7 – 8 – 9

Localidade B: 6 – 7 – 8 – 6 – 7 – 9 – 9 – 8 – 7 – 8

Localidade C: 7,5 – 6 – 6 – 6,5 – 7,5 – 6 – 5,5 – 8 – 8 – 7

Ao calcularmos as médias dessas três localidades, encontramos, respectivamente,

* Ronald Aylmer Fisher (1890 – 1962) foi um estatístico, biólogo evolutivo e geneticista inglês.

$\overline{X}_A = 7{,}5;$

$\overline{X}_B = 7{,}3;$

$\overline{X}_C = 6{,}8.$

O QUE DESEJAMOS SABER É SE ESSAS DIFERENÇAS ENTRE AS MÉDIAS SÃO SIGNIFICANTES OU SE PODEM SER ATRIBUÍDAS AO ACASO.

Chamaremos, então, de:

μ_A a média da localidade A,

μ_B a média da localidade B e

μ_C a média da localidade C.

Vamos testar a hipótese nula (H_0) que $\mu_A = \mu_B = \mu_C$.

A hipótese alternativa (H_1) afirma que as médias não são iguais (pelo menos duas são diferentes).

Se após o teste rejeitarmos H_0, concluímos que a propriedade considerada tem influência sobre a variável em análise. Lembrar que o teste leva em consideração determinado nível de significância (por exemplo, $\alpha = 5\%$).

O PRIMEIRO PASSO É CALCULARMOS AS MÉDIAS DE CADA AMOSTRA, a que chamaremos $\overline{X}_A$, $\overline{X}_B$ e $\overline{X}_C$, respectivamente.

Se $\overline{X}_A$, $\overline{X}_B$ e $\overline{X}_C$ estiverem próximas, tendemos a aceitar H_0 como verdadeira, ou seja, $\mu_A = \mu_B = \mu_C$.

Assim, O SEGUNDO PASSO É DETERMINAR A MÉDIA TOTAL (média das médias) pela fórmula:

$$\overline{X} = \frac{\overline{X}_A + \overline{X}_B + \overline{X}_C}{m} = \frac{7{,}5 + 7{,}3 + 6{,}8}{3} = 7{,}2$$

A variância da amostra, para as três médias, será obtida pela fórmula:

$$S^2 = \frac{(\overline{X}_A - \overline{X})^2 + (\overline{X}_B - \overline{X})^2 + (\overline{X}_C - \overline{X})^2}{m - 1}$$

$$S^2 = \frac{(7{,}5 - 7{,}2)^2 + (7{,}3 - 7{,}2)^2 + (6{,}8 - 7{,}2)^2}{2}$$

$$S^2 = \frac{0{,}09 + 0{,}01 + 0{,}25}{2} = 0{,}175$$

Calculamos, na sequência, a variância amostral para cada grupo de resultados.

X_A	X_B	X_C	$(X_A - \overline{X}_A)^2$	$(X_B - \overline{X}_B)^2$	$(X_C - \overline{X}_C)^2$
8	6	7,5	0,25	1,69	0,49
7	7	6	0,25	0,09	0,64
8	8	6	0,25	0,49	0,64
6	6	6,5	2,25	1,69	0,09
9	7	7,5	2,25	0,09	0,49
7	7	6	0,25	0,09	0,64
6	9	5,5	2,25	2,89	1,69
7	8	8	0,25	0,49	1,44
8	7	8	0,25	0,49	1,44
9	8	7	2,25	0,49	0,04
$\sum = 75$	$\sum = 73$	$\sum = 68$	$\sum = 10{,}5$	$\sum = 8{,}1$	$\sum = 7{,}6$

$$S_A^2 = \frac{\Sigma\,(X_A - \overline{X}_A)^2}{n-1} = \frac{10{,}5}{9} = 1{,}17$$

$$S_B^2 = \frac{\Sigma\,(X_B - \overline{X}_B)^2}{n-1} = \frac{8{,}1}{9} = 0{,}90$$

$$S_C^2 = \frac{\Sigma\,(X_C - \overline{X}_C)^2}{n-1} = \frac{7{,}6}{9} = 0{,}84$$

Então, a média dessas variâncias amostrais é igual a:

$$\overline{S^2} = \frac{S_A^2 + S_B^2 + S_C^2}{m} = \frac{1{,}17 + 0{,}90 + 0{,}84}{3} = 0{,}97$$

Quanto maiores forem S_A^2, S_B^2 e S_C^2, maior a probabilidade de $\overline{X}_A$, $\overline{X}_B$ e $\overline{X}_C$ serem diferentes. Logo, quanto maior for $\overline{S^2}$, maior a chance de aceitarmos H_0.

Agora que conhecemos a variância da amostra para as três médias (S^2) e a média das três variâncias amostrais ($\overline{S^2}$), podemos calcular a denominada *distribuição F* (em homenagem a Fisher), pela fórmula:

$$F = \frac{n \cdot S^2}{\overline{S^2}}$$

No nosso exemplo,

$$F = \frac{10 \,.\, 0{,}175}{0{,}97} = 1{,}80$$

Após a determinação do valor de F, precisamos consultar a tabela 37 para encontrar o valor crítico de uma distribuição F. Caso o valor observado seja maior que o valor crítico, rejeitaremos H_0 e, caso contrário, aceitaremos H_0.

Verifique na tabela que, como n = 10 (φ_2) e m= 3 (φ_1), temos, na tabela 37, o resultado 3,71 (esse é o valor crítico). Como F = 1,80 para o nosso exemplo, aceitaremos H_0.

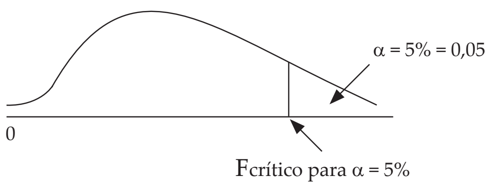
Tanto S^2 quanto $\overline{S}^2$ são estimativas não tendenciosas de σ^2 quando H_0 é verdadeira. Logo, são valores que tendem a ser iguais, o que faz com que F se aproxime de 1. Quando H_0 não é verdadeira, F assumirá um valor bem maior que 1.

> Resumindo, caso as médias sejam realmente iguais, F se aproxima de 1. Caso F seja muito maior que 1, rejeitaremos H_0.

Tabela 37 – Valores críticos de 5% de F

φ_2	Grau de liberdade do numerador de F (φ_1)																				
	1	2	3	4	5	6	7	8	9	10	12	14	15	16	18	20	24	30	40	60	120	∞
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	243,90	245,40	245,95	246,50	247,30	248,02	249,05	250,10	251,10	252,20	253,30	254,30
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,42	19,43	19,43	19,44	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,72	8,70	8,69	8,67	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,87	5,86	5,84	5,82	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,64	4,62	4,60	4,58	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,05	4,00	3,96	3,94	3,92	3,90	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,53	3,51	3,49	3,47	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,24	3,22	3,20	3,17	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,03	3,01	2,99	2,96	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,87	2,85	2,83	2,80	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,74	2,72	2,70	2,67	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,64	2,62	2,60	2,57	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2060	2,55	2,53	2,52	2,48	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,48	2,46	2,44	2,41	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,42	2,40	2,39	2,35	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,37	2,35	2,33	2,30	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,34	2,31	2,29	2,26	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,68	2,51	2,46	2,41	2,34	2,29	2,27	2,25	2,22	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,26	2,23	2,22	2,18	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,22	2,20	2,18	2,15	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,20	2,18	2,16	2,12	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,17	2,15	2,13	2,10	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,15	2,13	2,11	2,08	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,13	2,11	2,09	2,05	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,11	2,09	2,07	2,04	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,09	2,07	2,05	2,02	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,08	2,06	2,04	2,00	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,06	2,04	2,02	1,99	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,05	2,03	2,01	1,97	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,04	2,01	1,99	1,96	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,95	1,92	1,90	1,87	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,86	1,84	1,81	1,78	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,77	1,75	1,72	1,69	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,69	1,67	1,63	1,60	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Obs.: φ_2 é o grau de liberdade do denominador de F



Variância total é aquela que se obtém quando as m amostras são reunidas de modo a constituir uma única, composta da soma de todos os seus elementos.

Variância entre as amostras mede a variação existente entre todas as m amostras que são reunidas.

Variância dentro das amostras mede a variância dentro das n amostras tomadas em conjunto.

Se compararmos as médias de m amostras aleatórias de tamanho n, os graus de liberdade do denominador e do numerador são, respectivamente, m – 1 e m(n – 1).

Exercício Resolvido

Voltemos ao nosso exemplo anterior. Tínhamos m = 3 amostras e n= 10 elementos por amostra.

Então, temos, para a estatística F, 2 graus de liberdade para o numerador e 27 graus de liberdade para o denominador. Consultando a Tabela 37, teremos que a estatística F tem 95% de chance de ser menor do que 3,35. Resposta: Como o valor encontrado foi menor que 3,35, aceitaremos a hipótese nula (H_0).

Síntese

Como vimos, o método de análise da variância consiste em dividir a variância em componentes úteis. O modelo mais simples de Anova (*Analysis of Variance*) é aquele em que observamos os grupos considerando uma única propriedade, pois a aplicação desse método visa justamente possibilitar que façamos a comparação simultânea de várias médias de amostras. Nesse caso, denominamos a análise de distribuição de *F*, pois o método com essa finalidade foi criado por Fischer.

Questões para revisão

1. Quais os graus de liberdade para o numerador e para o denominador da distribuição F, considerando-se que compararemos as médias de 5 amostras aleatórias que possuem, cada uma, 12 elementos?

2. Um trabalhador de uma indústria produz determinada peça utilizando três diferentes máquinas. Produziu, em cada máquina, cinco peças e anotou o tempo gasto para cada uma, em segundos:

   Máquina A: 90 – 92 – 89 – 90 – 89

   Máquina B: 88 – 87 – 90 – 91 – 89

   Maquina C: 92 – 91 – 88 – 89 – 95

   Pergunta-se:

a) a hipótese nula poderá ser rejeitada?

b) sendo os dados fornecidos amostras aleatórias de populações normais e com o mesmo desvio padrão, as diferenças entre as três médias podem ser atribuídas ao acaso (faça o teste com 0,05 de nível de significância)?

3. Assinale as alternativas corretas.

São condições para a aplicação da análise da variância:

a. as amostras originarem-se em populações que têm distribuição anormal;

b. as amostras originarem-se em populações com distribuição normal;

c. o valor da variância ser o mesmo nas populações observadas;

d. a situação de dependência das amostras em relação a inferência feita.

4. Assinale (V) para as assertivas verdadeiras e (F) para as falsas.

( ) Constitui-se uma sequência de passos para a análise de variância: cálculo da média de cada amostra; determinação da média total; cálculo da variância amostral para cada grupo de resultados.

( ) A distribuição F é calculada após a média das variâncias amostrais.

( ) Usamos a distribuição F para verificarmos a igualdade entre duas médias.

( ) Fischer foi quem criou a Anova (*Analysis of Variance*) com o objetivo de estabelecer paralelos entre amostras aleatórias e independentes.

5. Assinale a alternativa correta. A variância pode ser:

a. apenas variância total.

b. unicamente a variância entre as amostras.

c. considerada sempre variância dentro das amostras.

d. variância total, variância entre as amostras e variância dentro das amostras.

Para concluir

Mário Quintana não era um estudioso de estatística, pelo menos não que se saiba. Mas foi de tal modo minucioso e atento que podemos até dizer ter ele muitas vezes perscrutado a estatística do insondável em seus poemas. Naturalmente, essa aplicação estatística pertence unicamente aos poetas. Mas que ele tinha uma consciência do fazer, da aplicação estatística, tinha. Veja o que ele afirmou em *As Indagações:* "A resposta certa, não importa nada: o essencial é que as perguntas estejam certas".

Sim, isso é primordial, ao aplicarmos os parâmetros estatísticos, precisamos saber objetivamente o que queremos medir. Esse é o princípio básico. Obviamente, como não somos poetas, as respostas devem conter uma margem de erro mínima. Aliás, considerando a aplicação da estatística no desenvolvimento das ciências, devemos destacar que não é suficiente observar os fenômenos, faz-se necessário o estudo de suas causas. É nesse processo que a estatística tornou-se uma ferramenta preciosa na pesquisa de casualidade que há entre os fenômenos. Atualmente, ela, a Estatística, apresenta-se como indispensável nos planejamentos, bem como na coleta, descrição, análise e organização das informações e dos conhecimentos. Assim, tornou-se uma ciência necessária em várias áreas das diversas ciências, e certamente em sua área de estudos ou atuação, não importa qual seja, ela também está presente, o que explica a importância de estudar a aplicação de métodos estatísticos para o seu aprimoramento profissional.

Referências por capítulo

Capítulo 1

[1] FERREIRA, Aurélio Buarque de Holanda. **Novo dicionário Aurélio**. 2. ed. Rio de Janeiro: Nova Fronteira, 1986.

[2] TOLEDO, Geraldo Luciano; OVALLE, Ivo Izidoro. **Estatística básica**. 2. ed. São Paulo: Atlas, 1995.

[3] VIEIRA, Sonia. **Princípios de estatística**. São Paulo: Pioneira, 1999.

[4] MARTINS, Gilberto de Andrade; DONAIRE, Denis. **Princípios de estatística.** 4. ed. São Paulo: Atlas, 1990.

[5] TOLEDO, Geraldo Luciano; OVALLE, Ivo Izidoro. **Estatística básica**. 2. ed. São Paulo: Atlas, 1995.

Capítulo 2

[1] TOLEDO, Geraldo Luciano; OVALLE, Ivo Izidoro. **Estatística básica**. 2. ed. São Paulo: Atlas, 1995.

[2] FONSECA, Jairo Simon da; MARTINS, Gilberto de Andrade. **Curso de estatística**. 6. ed. São Paulo: Atlas, 1996.

Capítulo 3

[1] WILD, Christopher J.; SEBER, George A. F. **Encontros com o acaso**: um primeiro curso de análise de dados e inferência. Rio de Janeiro: LTC, 2004.

Capítulo 4

[1] MARTINS, Gilberto de Andrade; DONAIRE, Denis. **Princípios de estatística**. 4. ed. São Paulo: Atlas, 1990.; FREUND, John E. **Estatística aplicada**: economia, administração e contabilidade. 11. ed. Porto Alegre: Bookman, 2006.

Capítulo 5

[1] LAPPONI, Juan Carlos. **Estatística usando Excel 5 e 7**. São Paulo: Lapponi Treinamento e Editora, 1997.

[2] FREUND, John E. **Estatística aplicada**: economia, administração e contabilidade. 11. ed. Porto Alegre: Bookman, 2006.

[3] TOLEDO, Geraldo Luciano; OVALLE, Ivo Izidoro. **Estatística básica**. 2. ed. São Paulo: Atlas, 1995.

Capítulo 6

[1] VIEIRA, Sonia. **Princípios de estatística**. São Paulo: Pioneira, 1999.

[2] TOLEDO, Geraldo Luciano; OVALLE, Ivo Izidoro. **Estatística básica**. 2. ed. São Paulo: Atlas, 1995.

[3] CASTRO, Lauro Sodré Viveiros de. **Pontos de estatística**. 16. ed. Rio de Janeiro: Científica, 1975.

Capítulo 7

[1] LIPSCHUTZ, Seymour. **Probabilidade**. Rio de Janeiro: McGraw-Hill do Brasil, 1974.

[2] CASTRO, Lauro Sodré Viveiros de. **Pontos de estatística**. 16. ed. Rio de Janeiro: Científica, 1975.

[3] FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. **Introdução à estatística**. 2. ed. Rio de Janeiro: LTC, 2003.

Capítulo 8

[1] HOEL, Paul Gerhard. **Estatística elementar**. São Paulo: Atlas, 1981.

Capítulo 9

[1] SILVA, Ermes Medeiros da. et al. **Estatística para os cursos de**: economia, administração, ciências contábeis. São Paulo: Atlas, 1997.

Capítulo 10

[1] FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. **Introdução à estatística**. 2. ed. Rio de Janeiro: LTC, 2003.

Capítulo 11

[1] DOWNING, Douglas; CLARK, Jeffrey. **Estatística aplicada**. São Paulo: Saraiva, 1998.

Capítulo 12

[1] BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica**. 5. ed. São Paulo: Saraiva, 2002.

[2] FONSECA, Jairo Simon da; MARTINS, Gilberto de Andrade. **Curso de estatística**. 6. ed. São Paulo: Atlas, 1996.

[3] HOEL, Paul Gerhard. **Estatística elementar**. São Paulo: Atlas, 1981.

Referências

BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica**. 5. ed. São Paulo: Saraiva, 2002.

CASTRO, Lauro Sodré Viveiros de. **Pontos de estatística**. 16. ed. Rio de Janeiro: Científica, 1975.

DOWNING, Douglas; CLARK, Jeffrey. **Estatística aplicada**. São Paulo: Saraiva, 1998.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. **Introdução à estatística**. 2. ed. Rio de Janeiro: LTC, 2003.

FERREIRA, Aurélio Buarque de Holanda. **Novo dicionário Aurélio**. 2. ed. Rio de Janeiro: Nova Fronteira, 1986.

FONSECA, Jairo Simon da; MARTINS, Gilberto de Andrade. **Curso de estatística**. 6. ed. São Paulo: Atlas, 1996.

FREUND, John E. **Estatística aplicada**: economia, administração e contabilidade. 11. ed. Porto Alegre: Bookman, 2006.

HOEL, Paul Gerhard. **Estatística elementar**. São Paulo: Atlas, 1981.

LAPPONI, Juan Carlos. **Estatística usando Excel 5 e 7**. São Paulo: Lapponi Treinamento e Editora, 1997.

LIPSCHUTZ, Seymour. **Probabilidade**. Rio de Janeiro: McGraw-Hill do Brasil, 1974.

MARTINS, Gilberto de Andrade; DONAIRE, Denis. **Princípios de estatística**. 4. ed. São Paulo: Atlas, 1990.

SILVA, Ermes Medeiros da. et. al. **Estatística para os cursos de**: economia, administração, ciências contábeis. São Paulo: Atlas, 1997.

TOLEDO, Geraldo Luciano; OVALLE, Ivo Izidoro. **Estatística básica**. 2. ed. São Paulo: Atlas, 1995.

VIEIRA, Sonia. **Princípios de estatística**. São Paulo: Pioneira, 1999.

WILD, Christopher J.; SEBER, George A. F. **Encontros com o acaso**: um primeiro curso de análise de dados e inferência. Rio de Janeiro: LTC, 2004.

Respostas

Capítulo 1

1. Resposta possível: População é o conjunto de elementos que desejamos observar para obter determinada informação.
2. Resposta possível: Amostra é o subconjunto de elementos retirados da população que se está observando.
3. b
4. c
5. F, F, F, V, F

Capítulo 2

1. 7
2. Cronológica
3. 5
4. Resposta possível: O gráfico aqui apresentado é o denominado *gráfico de colunas*.
5. Resposta possível: As partes que constituem uma tabela são o cabeçalho, o corpo e o rodapé.

Capítulo 3

1. 6
2. 60
3. 8
4. 2
5. 9
6. V, F, F, F, V
7. Resposta possível: Enquanto a variável qualitativa descreve qualidades, a variável quantitativa é expressa por meio de valores numéricos.

8. Resposta possível: Ela é uma série estatística específica caracterizada pelo fato de seus dados serem dispostos em classes, com suas respectivas frequências absolutas.

Capítulo 4

1. 7
2. R$ 830,40
3. 7
4. R$ 824,00
5. R$ 828,00
6. 3,0
7. 55
8. 15
9. 146
10. 440,00
11. Resposta possível: As medidas de posição servem para resumir os dados, apresentando um ou mais valores da série estudada.
12. Resposta possível: A média aritmética simples é utilizada para dados não agrupados e a média aritmética ponderada para dados agrupados, onde cada grandeza envolvida no cálculo da média tem diferente peso.

Capítulo 5

1. 2
2. 5, 6
3. 2,3664
4. 6
5. 2
6. 1,76
7. 4,33
8. 2,08

9. Resposta possível: As medidas de dispersão são utilizadas para verificar o quanto os valores encontrados em uma pesquisa estão afastados ou dispersos em relação à média ou a em relação à mediana.
10. Resposta possível: As medidas de dispersão são: amplitude total ou intervalo total; amplitude semi-interquartílica, ou intervalo semi-interquartílico, ou desvio quartil; desvio médio; variância; e desvio padrão.

Capítulo 6

1. – 0,20
2. 0,30
3. platicúrtica, com K = 0,333
4. platicúrtica
5. simétrica
6. Resposta possível: Elas indicam o grau de deformação de uma curva de frequências.
7. Resposta possível: Elas indicam o quanto uma distribuição de frequências é mais achatada ou mais afilada do que uma curva padrão, a qual é denominada de *curva normal*.

Capítulo 7

1. 10/18
2. 1/2
3. 4/16
4. 16/52
5. 55/100
6. 76%
7. 5%
8. 284/3000
9. 8/720
10. 9/20
11. 38/100

12. 0,16%
13. Resposta possível: A estatística baseia-se em experimentos. O cálculo de probabilidades, por sua vez, baseia-se em postulados lógicos.
14. Resposta possível: Os fenômenos estudados em estatística são fenômenos cujos resultados apresentam variações de uma observação para outra, dificultando dessa maneira a previsão de um resultado futuro.

Capítulo 8

1. 43,05%
2. 19,37%
3. 14,68%
4. 9,375%
5. 26,68%
6. Resposta possível: É um modelo matemático para a distribuição real de frequências.
7. Resposta possível: Na variável aleatória, os valores são determinados por processos ocasionais, aqueles que ocorrem fora do nosso controle como observadores. Ela pode ser discreta ou contínua. Quando se trata de uma variável aleatória discreta, podemos relacionar todos os possíveis valores em uma tabela com as respectivas probabilidades. Quando for uma variável contínua, não podemos listar todos os possíveis valores fracionários e teremos que retratar as probabilidades por meio de uma função densidade ou por uma curva de probabilidade.

Capítulo 9

1. 27,068%
2. 6,13%
3. 2,90%
4. 14,66%
5. 4,46%

6. Resposta possível: Utilizamos a distribuição de Poisson para determinar a probabilidade de um dado número de sucessos quando os eventos ocorrem em um contínuum de tempo ou espaço.
7. Resposta possível: Somente um. Ou seja, o número médio de sucessos para a específica dimensão de interesse.

Capítulo 10

1. 20 alunos
2. 68,26%
3. 15,87%
4. 77,45%
5. 92,70%
6. 2,28%
7. 30,85%
8. 5,45
9. 93,32%
10. 31,74%
11. Resposta possível: Uma variável aleatória contínua caracteriza-se pelo fato de possuir a condição de assumir qualquer valor real (inteiro ou fracionário) dentro de um intervalo definido de valores.
12. Resposta possível: Ela caracteriza-se como uma distribuição de probabilidade contínua. O seu aspecto diferencial está no fato de ser simétrica em relação à média e mesocúrtica e assíntota em relação ao eixo das abscissas, em ambas as direções.

Capítulo 11

1. $\bar{X} = n = 12$

   $S^2 = 2n = 24$

   $S = 4{,}89898$
2. $k = 13{,}675$
3. $k = 0{,}975 - 0{,}025 = 0{,}950$ ou 95%

4. k = 9,299
5. k = 56,369
6. b
7. b

Capítulo 12

1. IC (67,38 < μ < 68,62) = 90%.
2. IC (159 < μ < 165) = 95%.
3. IC (56,29 < μ < 70,25) = 99%.
4. IC (60,10 < μ < 66,44) = 75,80%.
5. IC (1780 < μ < 1900) = 95%.
6. Resposta possível: A inferência estatística compreende uma sequência de procedimentos probabilísticos, em que estão inclusos o de amostragem, estimação e aplicação do intervalo de confiança.
7. Resposta possível: São valores obtidos através de observações de uma amostra que utilizamos para determinar um intervalo de modo que haja a probabilidade de tal intervalo conter o valor desconhecido de um parâmetro que buscamos.

Capítulo 13

1. $z_r = -2{,}67$ e está na zona de rejeição.
2. $z_r = 1{,}5$ e está na zona de aceitação.
3. A hipótese que μ = 500 é aceita, pois $z_r < 1{,}65$.
4. A hipótese nula será aceita porque z_r está na zona de aceitação.
5. A hipótese nula será aceita porque z_r está na zona de aceitação.
6. Resposta possível: Hipótese nula é justamente a informação ou hipótese que será testada (H_O), enquanto a hipótese alternativa é aquela que afirma ser a hipótese nula falsa. A primeira (a nula) representa uma igualdade e a segunda (a alternativa) uma desigualdade.

7. Resposta possível: Recebe o nome de *nível de significância* a probabilidade de cometermos o erro de rejeitar a hipótese nula quando ela for verdadeira.

Capítulo 14

1. Numerador = grau de liberdade igual a 55.
   Denominador = grau de liberdade igual a 4.

2) a) Não, pois F = 1,11.
   b) Sim, podem ser atribuídas ao acaso.

3. b, c

4. V, V, F, F

5. d

Sobre o autor

Nelson Pereira Castanheira é graduado em Eletrônica pela Universidade Federal do Paraná – UFPR e em Matemática, Física e Desenho Geométrico pela Pontifícia Universidade Católica do Paraná – PUCPR (1976), especialista em Análise de Sistemas, Finanças e Informatização, mestre em Administração de Empresas com ênfase em Recursos Humanos e doutorado em Engenharia da Produção pela Universidade Federal de Santa Catarina – UFSC.

Ao longo de sua carreira, exerceu diversas atividades: professor e coordenador da Escola Técnica Federal do Paraná; gerente de produtos e serviços da Telebahia; instrutor e analista de dados da Telepar; professor da Uniandrade em Curitiba-PR; professor assessor da Coordenação do Curso de Administração da Universidade Tuiuti do Paraná. São 35 anos de experiência no magistério e, em parelelo, 30 anos na área empresarial.

Atualmente, é coordenador de curso na Faculdade de Tecnologia Internacional – Fatec Internacional e professor do Instituto Brasileiro de Pós-Graduação e Extensão – Ibpex. Já tem oito obras publicadas.

Esta obra oferece farto conteúdo teórico e prático e constitui-se em uma possibilidade de estudo que satisfaz às necessidades didáticas tanto do ensino presencial como do ensino a distância pela forma lógica com que os assuntos são abordados. Os vários tópicos – como distribuição de frequências, medidas de tendência central, medidas de dispersão, assimetria, curtose, teoria das probabilidades, distribuição de probabilidades, estimação, intervalos de confiança e teste de hipóteses – são desenvolvidos com exemplos e exercícios para facilitar a compreensão e tornar o processo de aprendizagem mais eficiente e prazeroso.

Trata-se de um material prático que possibilita o entendimento e o uso desta tecnologia quantitativa, que é instrumento valioso para a tomada de decisões de profissionais das mais variadas áreas de atuação que recorrem à análise de dados para subsidiar suas atividades cotidianas.

EDITORA IBPEX DIALÓGICA
www.editoraibpex.com.br

ISBN 978-85-7838-943-7
9 788578 389437